Final Project

Online Shoppers Purchasing Intention

DATSKUYY - Data Scientist Team in E-Grocery **“Lauk Fresh”**

M. **Hendrawan** H.

Jimmy Firdaut

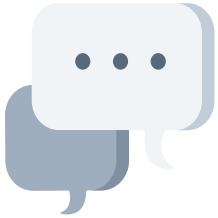
Erlando Febrian

M **Arfanul** Aziz

Hafizh Adi Prasetya
*Our **Mentor***

Salman Al Farisi

Table of Contents

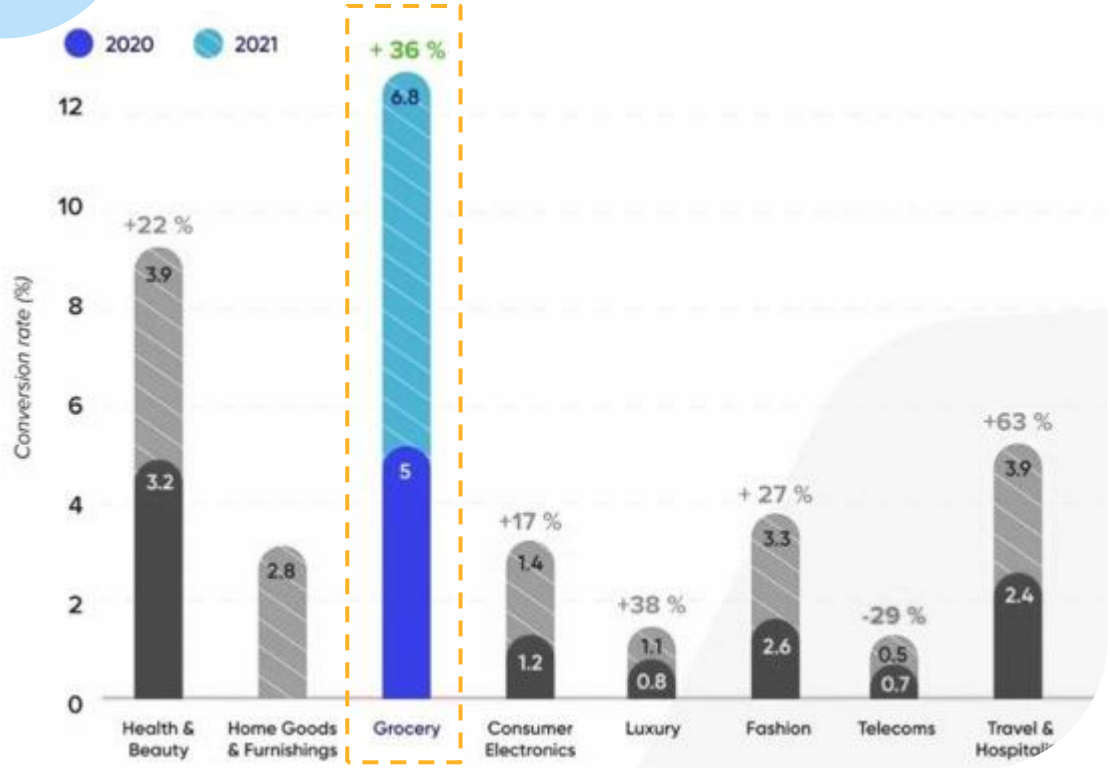


01. Project Background

"Turnover is vanity,
profit is sanity"

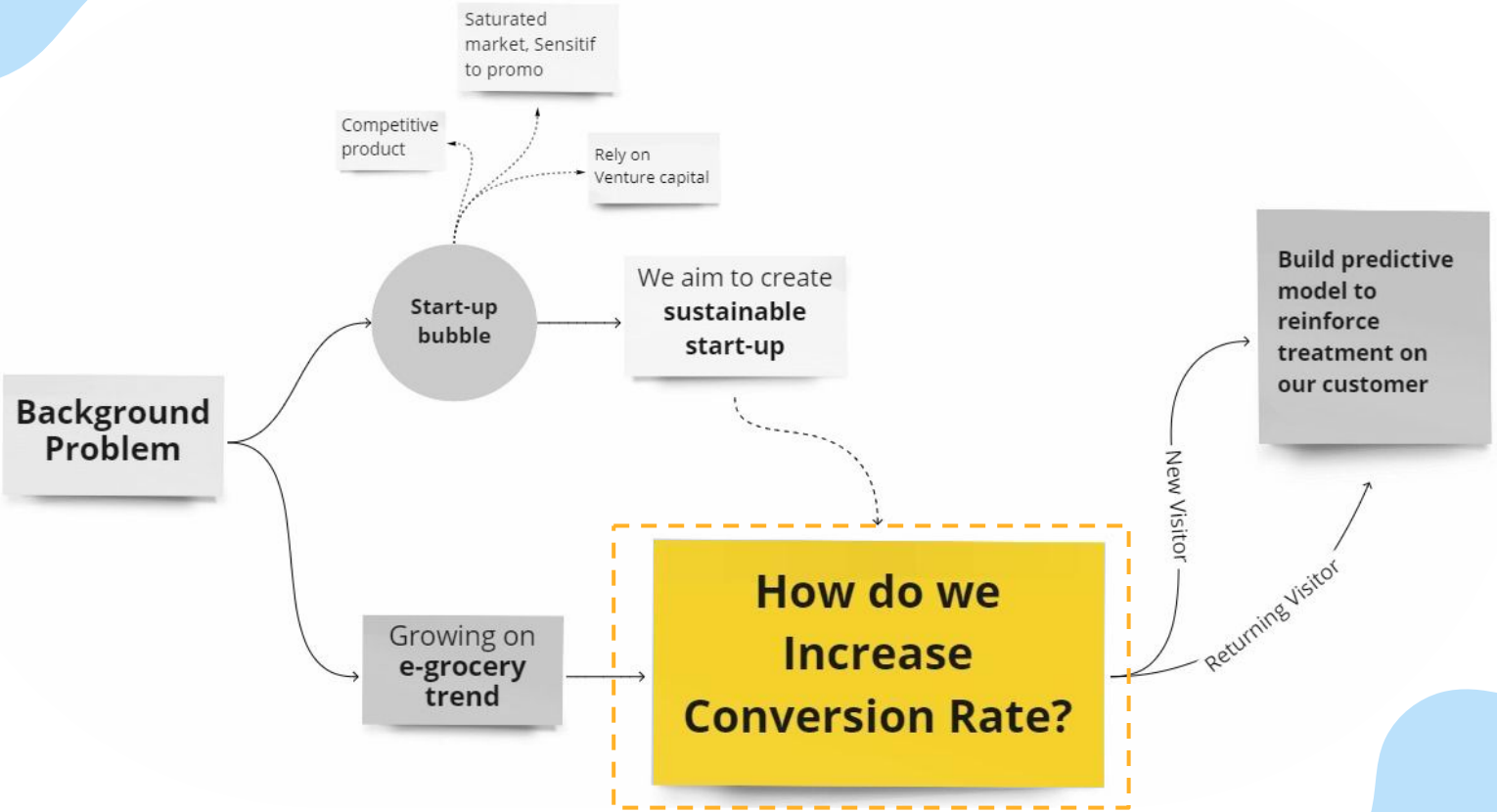
– Wise Businessman –

Average Conversion rate for eCommerce Industries



The data clearly shows the impact of the pandemic: **Grocery has the highest increase in conversion rate (+36%) in 2021 vs 2020**, as many consumers have permanently shifted to buying their groceries online.

Background


Saturated market, Sensitif to promo
Competitive product
Rely on Venture capital
Start-up bubble
We aim to create sustainable start-up
Background Problem
Growing on e-grocery trend
How do we Increase Conversion Rate?
New Visitor
Returning Visitor
Build predictive model to reinforce treatment on our customer

Our Objectives

Goal

Increase the company's conversion rate up to **20%**

and gross profit up to **15%** in 2023

Metric

- Conversion rate
- Gross Profit

Objective

- Analyze the factors that affect the increase in conversion rate
- Predict whether visitors will convert or not using predictive modeling

02. Exploratory Data Analysis

Dataset Overview

1 Year Historical Data (12,330 Sessions)

Numerical Features

- Administrative
- Administrative_Duration
- Informational
- Informational_Duration
- ProductRelated
- ProductRelated_Duration
- BounceRates
- ExitRates
- Page Values
- Special Day

Categorical Features

- Month
- OperatingSystems
- Browser
- Region
- Traffic_Type
- Visitor_Type
- Weekend

Target (Imbalanced)

- Purchase

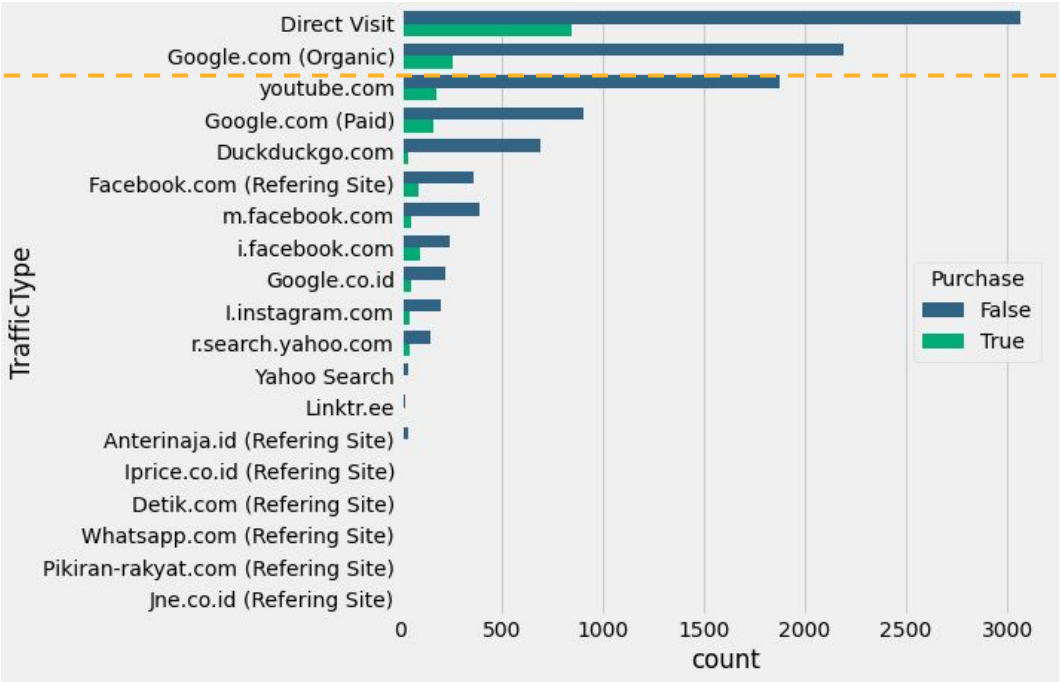
Feature Imputation

We need to do imputation for numeric code in
Categorical Features into more understandable values

1. **OperatingSystems**[11]
   - ‘1’ to ‘Android 11’
   - ‘3’ to ‘Win 10’
   - etc.
2. **Browser**[12]
   - ‘2’ to ‘Chrome Mobile’
   - ‘13’ to ‘Safari Desktop’
   - etc.
3. **Region**[13]
   - ‘4’ to ‘Jakarta’
   - ‘3’ to ‘Depok’
   - etc.
4. **Traffic_Type**[14]
   - ‘2’ to ‘Direct Visit’
   - ‘1’ to ‘Youtube.com’
   - etc.

Exploratory Data Analysis

(TrafficType Feature)

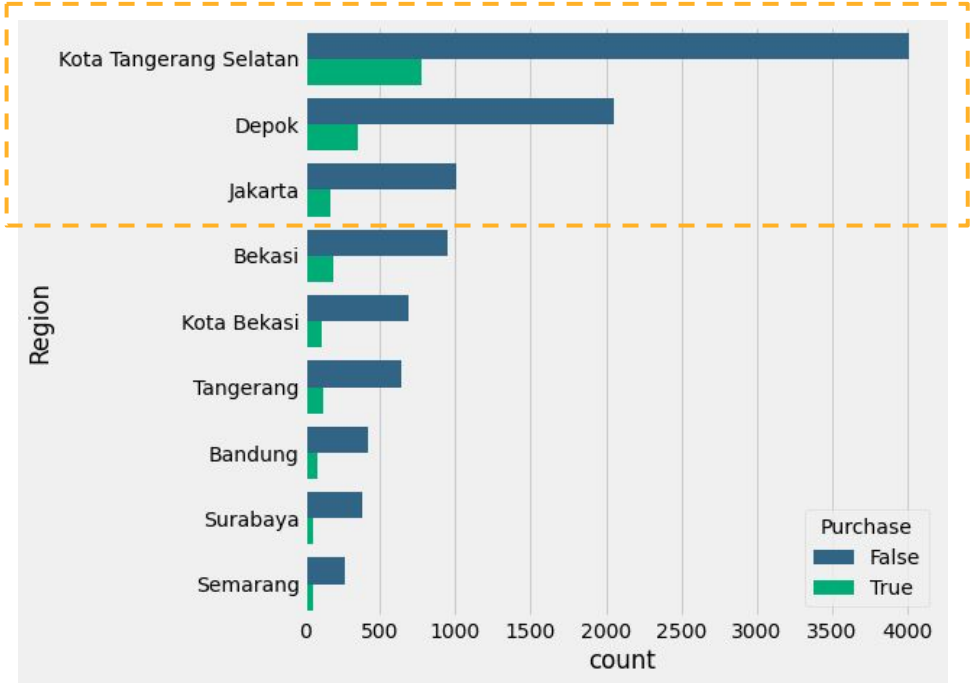


Most traffic come from organic traffic. **847 visitors make purchase through Direct Visit** and 3066 didn't. **262 Visitors make purchase through Google** Search Engine and 2189 didn't.

**) Traffic source by which the visitor has arrived at the Website/Platform*

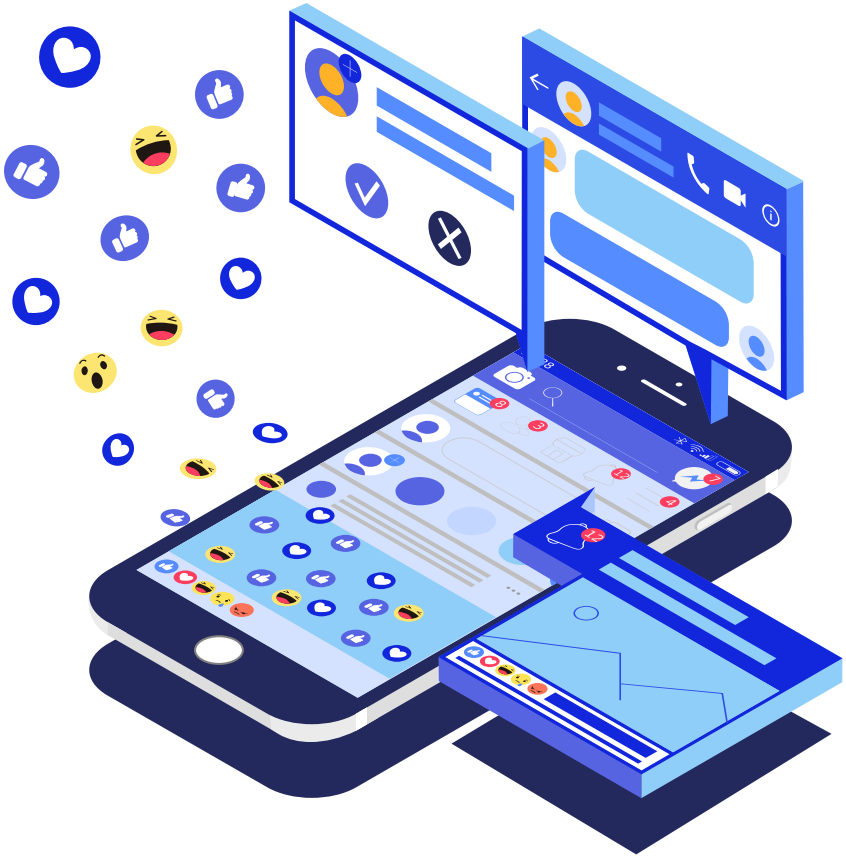
Exploratory Data Analysis

(Region Feature)



Most of visitors come from the Greater Jakarta area, with the top 3: Kota Tangerang Selatan, Depok, and Jakarta**. 771 visitors from Tangerang Selatan make purchase,** 4009 didn't. **349 visitors from Depok make purchase,** 2054 didn't.

**) Geographic region from which the session has been started by the visitor*



03. Data Pre Processing

Data Pre Processing

Handling Duplicate

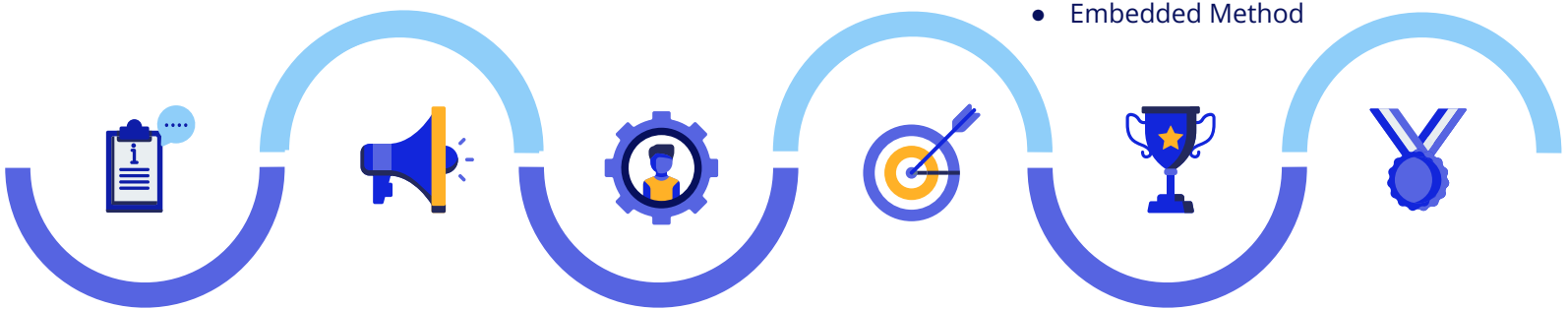
Drop 125 Duplicated Rows

Data Transformation

- Log Transformation
- MinMaxScaler

Feature Selection

- Drop Irrelevant Features
- Filter Method
- Embedded Method



Feature Encoding

One Hot Encoding

Handling Outlier

Z-score Method

Handling Imbalance Target

SMOTE 1:1

4. Modelling

Classification Model

	Train Precision	Test Precision	Train Recall	Test Recall	Train F2 Score	Test F2 Score
LogisticRegression	0.69	0.92	0.65	0.65	0.66	0.69
KNeighborsClassifier	0.95	0.88	0.77	0.68	0.8	0.71
GaussianNB	0.68	0.91	0.63	0.63	0.64	0.67
SVC	0.78	0.91	0.67	0.65	0.69	0.69
DecisionTreeClassifier	1	0.86	1	0.77	1	0.79
RandomForestClassifier	1	0.88	1	0.86	1	0.86
AdaBoostClassifier	0.77	0.92	0.67	0.66	0.69	0.7
GradientBoostingClassifier	0.83	0.9	0.79	0.79	0.8	0.81
LGBM Classifier	0.89	0.87	0.94	0.9	0.93	0.89
XGBClassifier	0.83	0.9	0.79	0.78	0.8	0.8

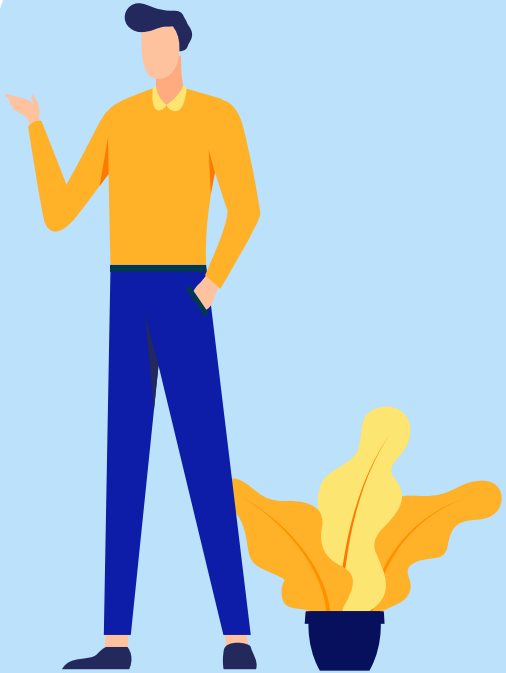
*) Positive Label = 0

Hyperparameter Tuning

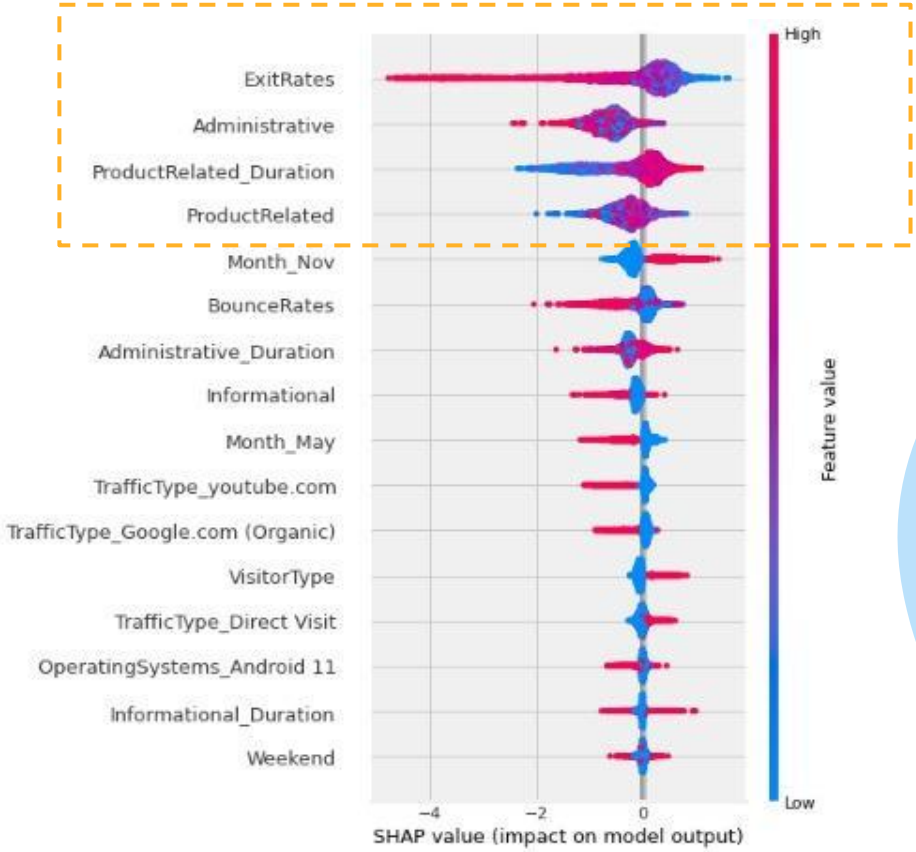
Algorithm	Train Precision	Test Precision	Train Recall	Test Recall	Train F2-Score	Test F2-Score
LGBM Classifier	93%	86%	97%	91%	96.4%	90.1%
Random Forest Classifier	100%	88%	100%	86%	100%	86.2%

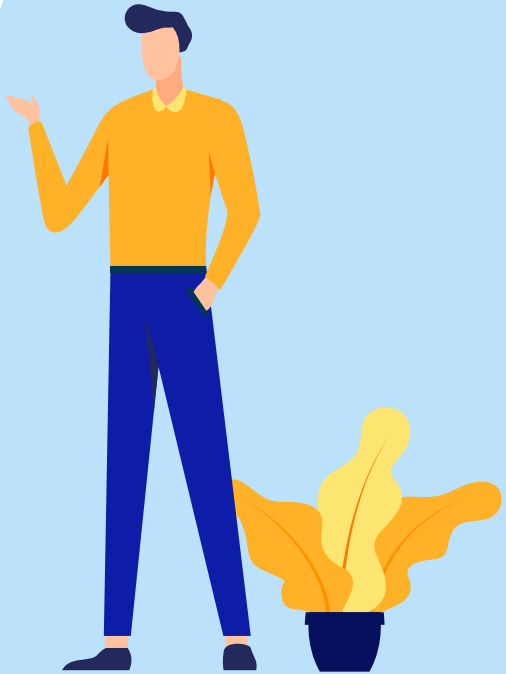
We decide to Interprate **LGBM Classifier Model** and extract its Feature Importance with SHAP, because :

- Has high F2 Score
- Goodfit



Feature Importance

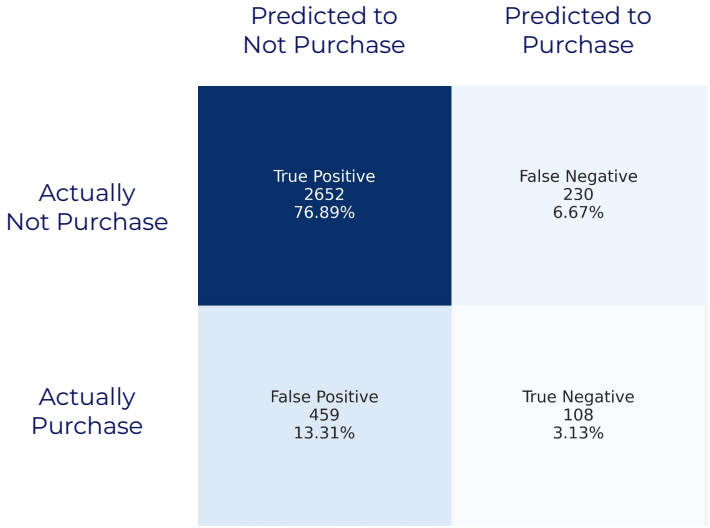




Final Model

**) Only use Top 4 Features Importance*

Algorithm	Train Precision	Test Precision	Train Recall	Test Recall	Train F2-Score	Test F2-Score
LGBM Classifier	86%	85%	97%	92%	94.3%	90.6%



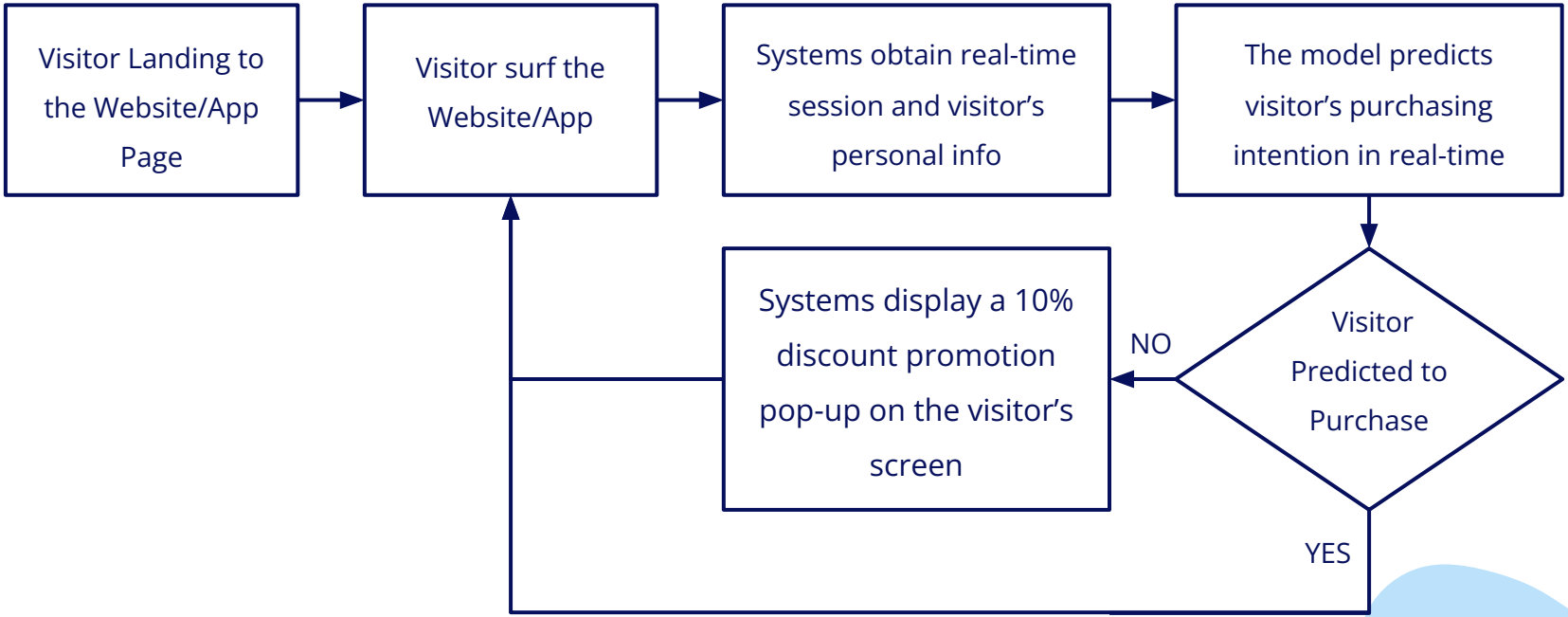
We decide to use this model as a **Final Model**, because :

- Has highest F2 Score
- Less overfit than before (use all features)

Business Insights & Recommendation

Discount Offering to Real-Time Non-Purchase Predicted Visitors

First Recommendation Workflow



Discount Offering to Real-Time Non-Purchase Predicted Visitors

First Recommendation Simulation

BEFORE		***AFTER***
15.63% Conversion Rate / Year	➡	**23.32%** Conversion Rate / Year
1.01B Gross Profit / Year	➡	**1.17B** Gross Profit / Year

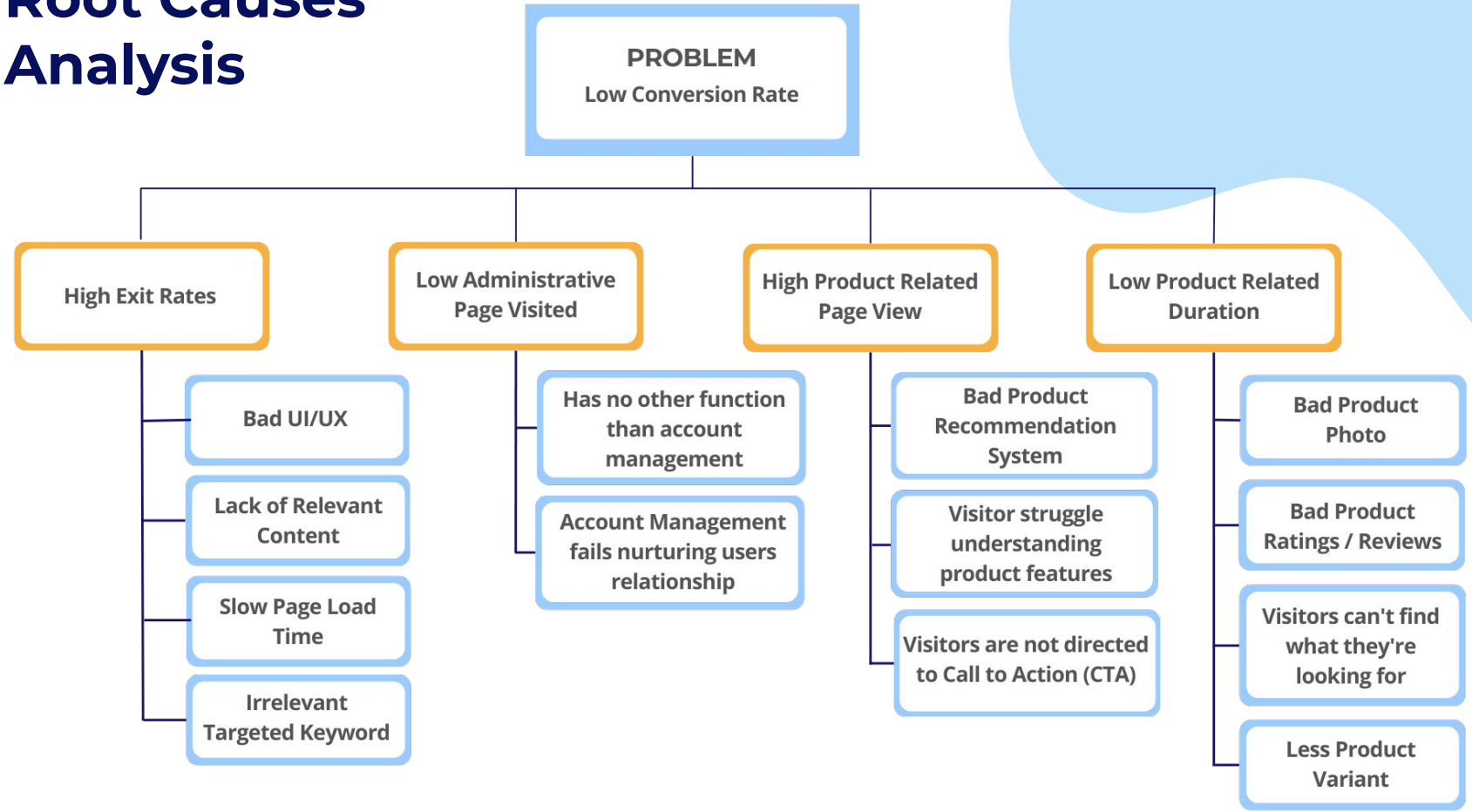
**) Gross Profit / Year = (Revenue - COGS) x Visitors / Year*

Assumption :

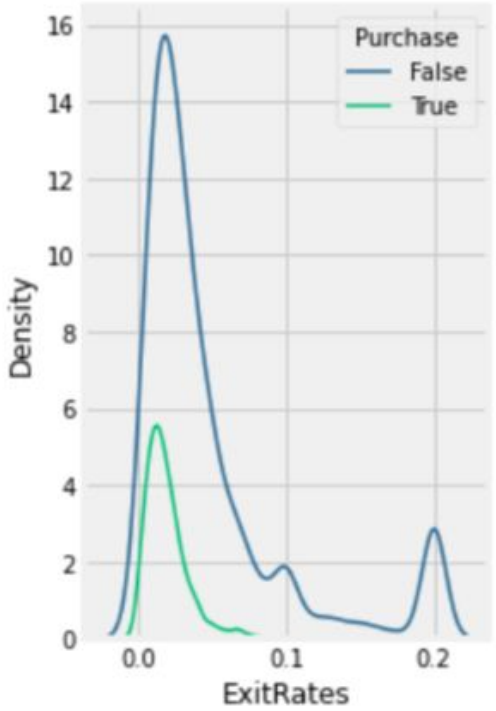
1. *COGS avg / purchase = Rp100.000*
2. *Gross Profit Margin = 65%*
3. *Promotion Effective Rate = 10%*
4. *# Visitors / year = 100.000*

First Recommendation potentially increase **7.69% Conversion Rate** and **17% Gross Profit**.

Root Causes Analysis



Business Insights (Exit Rates Feature)



Visitors who don't make a purchase have higher ExitRates, and it's known that the median Exit Rates for people who purchase are **1.6%** and those who do not purchase are **2.8%.**

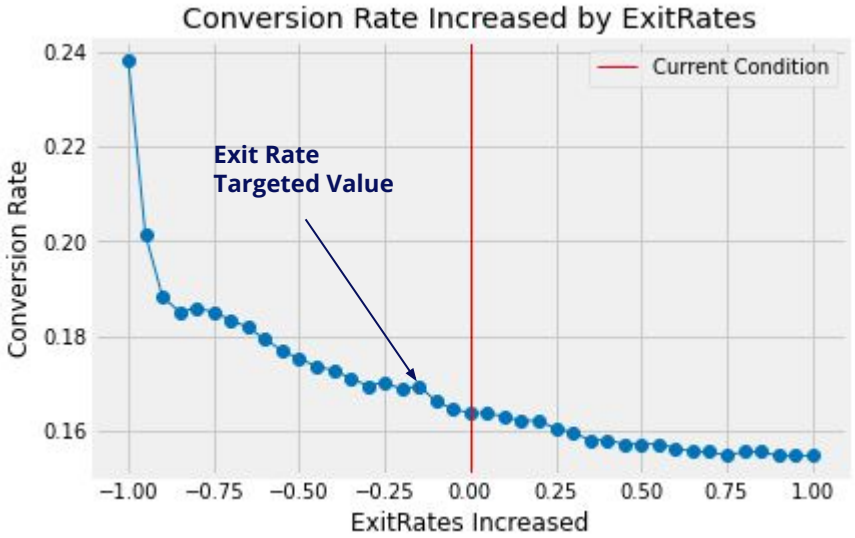
Insights:

- Need to find the common causes why our platform have a high Exit Rates
- Provide several best recommendations to cope with high exit rates

**) Exit Rate: Average exit rate value of the pages visited by the visitor*

Exit Rate Sensitivity Analysis

Second Recommendation Simulation (Lowering Exit Rates)



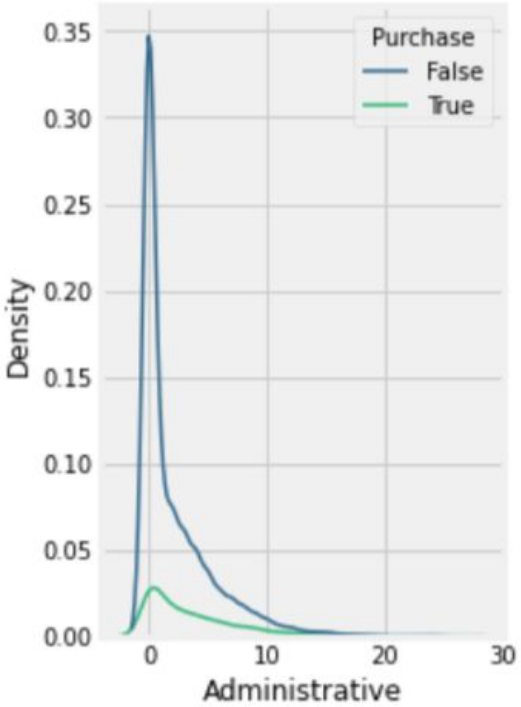
Based on the experiments result, we are looking for the optimal reduction in Exit Rates to increase our Conversion rate, we decide to **Decrease 15% Exit Rates.** By implementing this, targeted Exit Rate is expected to **increase the conversion rate up to 0.6%**

Exit Rate Improvement

Second Recommendation (Lowering Exit Rates)

Root Cause	Recommendation	How to do	Pros	Cons
Bad UI/UX[1]	Improve UI/UX Design	• Conduct Through Research • Simplicity is a Must • Experimental Design	• Have a long-term effect • Relatively cheap	• Requires long working time • Need additional working time to do A/B testing
Lack of Relevant Contents[2]	Content Quality Improvement[5]	• Research and testing regularly • Reuse best contents • Track visitor activities in detail (Use heatmap software)	• Relatively cheap • Need less resources (time & cost) sthan improve UX	• Must be monitored regularly • Often depends on the current trends
Slow Page Loading Time[3]	Improve Page Loading Time[8]	• Upgrade web hosting • Optimize image (compression) • Upgrade CMS and it's plugin	Potentially increase conversion rate higher than other recommendation	• Requires good hosting and it's quite expensive • Need high technical skill team and it is costsly to get qualified employees
Irrelevant Targeted Keywords	SEO Improvement[6]	• Create content based on keyword[7] • Optimize on page SEO • Go after featured snippets	• Relatively cheap • Have a long term effect	• Need regularly improvement • It needs longer time to work on than other recommendation

Business Insights (Administrative Feature)



Administrative Pages shows positively skewed. Median value for visitors who do purchase are **2 pages** and those who don't purchase are **0 page**.

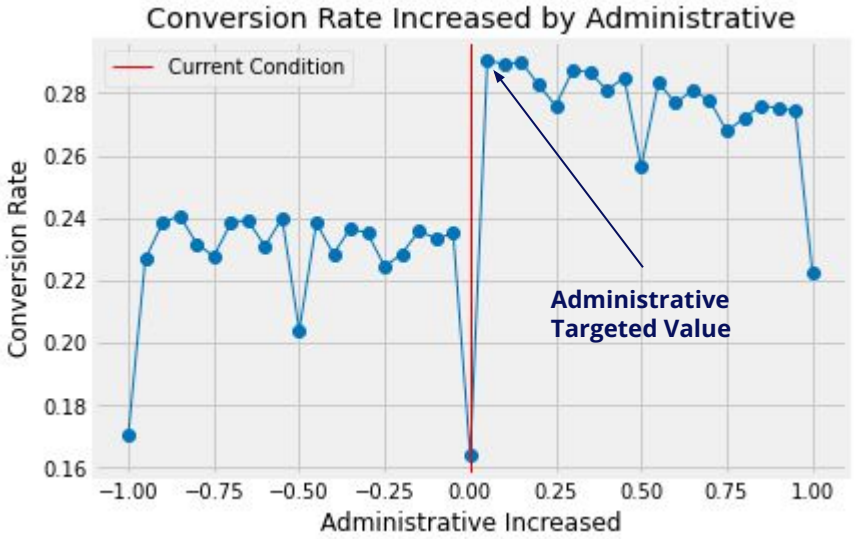
Insights:

- Need further investigation on this feature to get complete understanding
- Since this feature dealing with account management, it seems we need to add some exciting content/product on this page, so that the value of this feature will increase as well as engagement of the visitors

**) Administrative: Number of pages visited by the visitor about account management*

Administrative Sensitivity Analysis

Third Recommendation Simulation (Increase Num of Administrative Pages Visits)



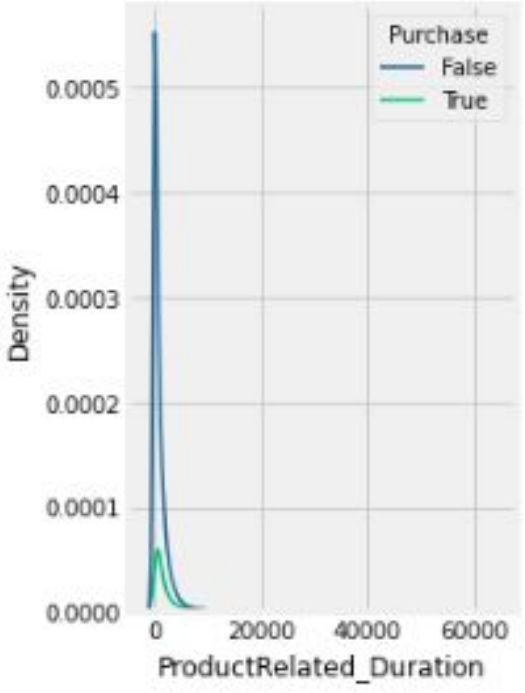
Based on the experiments result, we are looking for the optimal Administrative increment to increase our conversion rate, we decide to **Increase 5% Administrative.** By implementing this, targeted Administrative is expected to **increase the conversion rate up to 12.7%**

Administrative Improvement

Third Recommendation (Increase Num of Administrative Pages Visits)

Root Cause	Recommendation	How to do	Pros	Cons
Has no other function than account management (Too basic) Account Management fails nurturing user relationships	Improve UI/UX Design[4] Or Launch new features	• Increase visitor engagement by providing useful/exiting content or features (eg: Member point from ecommerce, gamification) • Do research and testing to get the best solution for this problem	Potentially increase administrative page views as visitor engagement improve	• Need longer time to do • Need additional time / resources to do A/B testing

Business Insights (Product Related Duration)



Visitors who don't make a purchase have enormous 0 second Product Related Duration, known that the median Product Related Duration for people who purchase is **8,5 min** and those who do not purchase are **18,4 min.**

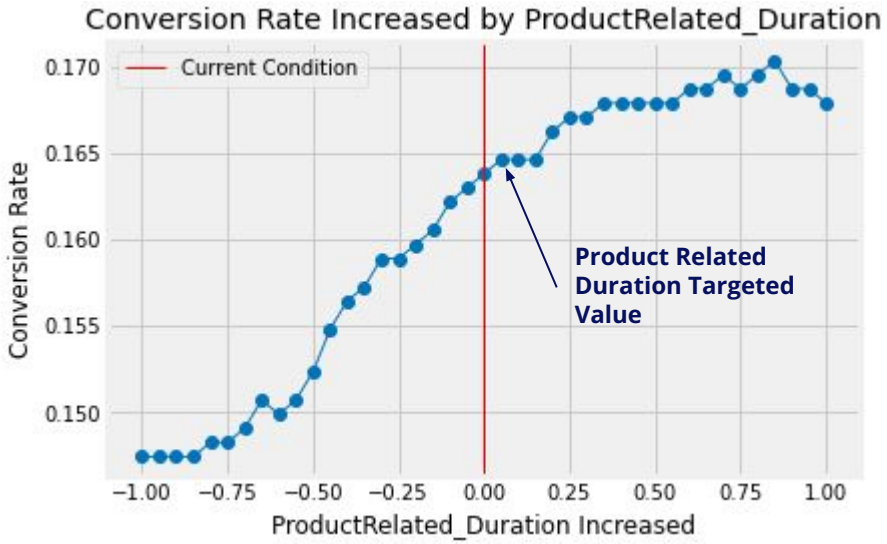
Insights:

- Need to find the common causes why our platform have a high Exit Rates
- Provide several best recommendations to cope with high Product Related Duration

**) ProductRelated_Duration: Total amount of time (in seconds) spent by the visitor on product related pages*

Product Related Duration Sensitivity Analysis

Fourth Recommendation Simulation (Increase Num of Duration People who Visit Product Related Pages)



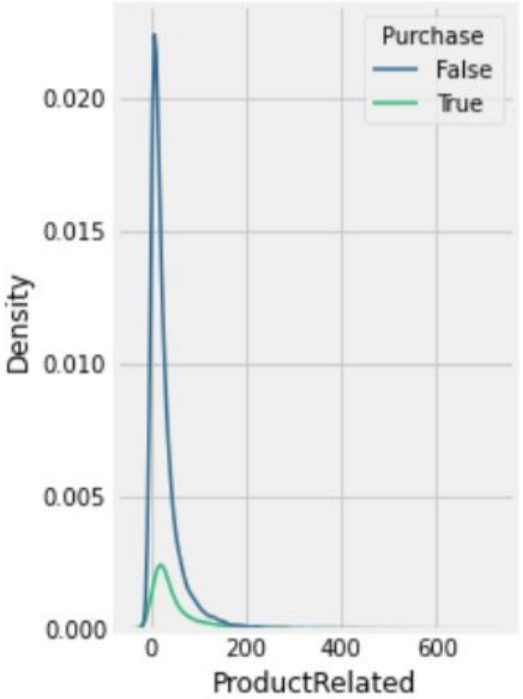
Based on the experiments result, we are looking for the optimal Product Related Duration increment to increase our conversion rate, we decide to **Increase 5% Product Related Duration.** By implementing this, targeted Product Related Duration is expected to **increase the conversion rate up to 0.6%**

Product Related Duration Improvement

Fourth Recommendation (Increase Num of Duration People Visit Product Related Pages)

Root Cause	Recommendation	How to do	Pros	Cons
Bad Product Recommendation System	Improve product recommender system	By improving the product recommender system, visitors can be more focused on the products they want and its variation, through this way, it can increase the Product Related Duration.	This method is very effective compared to other recommendations	• Need additional cost • Need qualified employees to do this work
Visitor struggle understanding product features	Improve platform's layout	• Conduct Research • Testing	• Has a long term effect • Relatively cheap	• Requires longer work time • Need additional work time to do A/B Testing

Business Insights (Product Related)



Product Related value **mostly below 100 pages**. known that the median **Product Related** for people who purchase are **29** pages and those who do not purchase are **16 pages**.

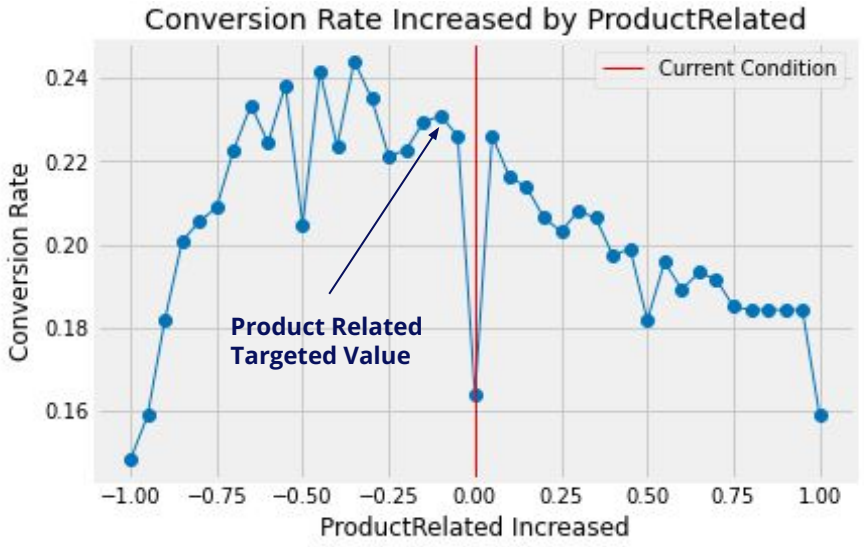
Insights:

- Need further investigation what causes this phenomenon
- Then establish some possible business recommendations based on finding

**) Product Related: Number of pages visited by visitor about product related pages*

Product Related Sensitivity Analysis

Sixth Recommendation Simulation (Decrease Num of People Visit Product Related Page)



Based on the experiments result, we are looking for the optimal Product Related reduction to increase our conversion rate, we decide to **Decrease 10% Product Related**. By implementing this, targeted Product Related is expected to **increase the conversion rate up to 6.7%**

Product Related Improvement

Fifth Recommendation (Increase Num of Duration People Visit Product Related Page)

Root Cause	Recommendation	How to do	Pros	Cons
Bad Photo Product	Product Photo Quality Improvement	Establish standard product display	• Whether using internal team / professional agency, it's quite cheap to afford • Direct impact since product photo is ready to publish	High quality photo will affect to page speed
Bad Product Ratings / Reviews	Improve product ratings and review[9]	• Conduct analysis to find the real causes of bad ratings and negative reviews • Solve the problem based on the main cause (eg: product quality, platform layout, etc)[10]	• Good reviews and ratings make a good social proof about our brand • Good product reviews also encourage new visitor to become a loyal visitor and might come back later (as returning visitor)	• Need effort to follow-up each of product which has bad reviews • Hard to find what specifications that customer really wants

Combine All Recommendation

Last Recommendation Simulation

BEFORE		***AFTER***
15.63% Conversion Rate / Year	➡	**30.40%** Conversion Rate / Year
1.01B Gross Profit / Year	➡	**1.96B** Gross Profit / Year

**) Gross Profit / Year = (Revenue - COGS) x Visitors / Year*

**) Assumption same as before*

Implemented Recommendation :

1. *Discount Offering to Real-Time Non-Purchase Predicted Visitors*
2. *Decrease **15% Exit Rates***
3. *Increase **5% Administrative***
4. *Increase **5% Product Related Duration***
5. *Decrease **10% Product Related***

With combine all recommendations potentially increase **14.77% Conversion Rate** and **95% Gross Profit**.

THANK YOU

APPENDIX

Features Dictionary (Numeric)

1. **Administrative** - Number of pages visited by the visitor about account management
2. **Administrative duration** - Total amount of time (in seconds) spent by the visitor on account management related pages
3. **Informational** - Number of pages visited by the visitor about Web site, communication and address information of the shopping site
4. **Informational duration** - Total amount of time (in seconds) spent by the visitor on informational pages
5. **Product related** - Number of pages visited by visitor about product related pages
6. **Product related duration** - Total amount of time (in seconds) spent by the visitor on product related pages
7. **Bounce rate** - Average bounce rate value of the pages visited by the visitor
8. **Exit rate** - Average exit rate value of the pages visited by the visitor
9. **Page Value** - Average page value of the pages visited by the visitor
10. **Special day** - Closeness of the site visiting time to a special day

Features Dictionary (Category)

1. **OperatingSystems** - Operating system of the visitor
2. **Browser** - Browser of the visitor
3. **Region** - Geographic region from which the session has been started by the visitor
4. **TrafficType** - Traffic source by which the visitor has arrived at the Web site / Platform
5. **VisitorType** - Visitor type as “New Visitor,” “Returning Visitor,” and “Other”
6. **Weekend** - Boolean value indicating whether the date of the visit is weekend
7. **Month** - Month value of the visit date
8. **Revenue (Purchase)** - Class label indicating whether the visit has been finalized with a transaction

References

1. https://databox.com/lower-exit-rate#number
2. https://unbounce.com/conversion-rate-optimization/high-bounce-rates/
3. https://unbounce.com/conversion-rate-optimization/high-bounce-rates/
4. https://www.google.com/url?q=https://bizeducator.com/10-benefits-and-importance-of-ux-design/&sa=D&source=docs&ust=1656576670066302&usg=AOvVaw2JNpY_F9OEMOycc4p4Is_N
5. https://www.searchenginejournal.com/tips-tricks-improve-content/294633/#close
6. https://www.semrush.com/blog/improve-seo/
7. https://www.crazyegg.com/blog/benefits/
8. https://designwebkit.com/web-and-trends/slash-ecommerce-website-load-time/9
9. https://optinmonster.com/get-more-product-reviews/
10. https://www.google.com/url?q=https://magenest.com/en/product-rating/&sa=D&source=docs&ust=1656576979331856&usg=AOvVaw0dxA4jRucHXPGvFZ6o9-rO

References

11. https://gs.statcounter.com/
12. https://gs.statcounter.com/
13. Google Trends
14. Tool SEO SEM rush

Pre Processing

1. **Drop Duplicated Rows**
   125 Rows Dropped

2. **Feature Encoding**
   a. Encoding result : 41 new features
   b. Drop original categorical features : Drop 5 features
   c. Total Features : 52 features

3. **Data Transformation**
   a. Log Transformation
   b. Scaler: MinMaxScaler

4. **Outlier Handling**
   Z-score method
   a. Before outlier handling : 12205 rows
   b. After outlier handling: 11495 rows

Pre Processing

5. **Train Test Split**
   Test size = 0.3
   a. Total train data : 8046 rows
   b. Total test data : 3449 rows

6. **Feature Selection**
   a. Drop Irrelevant Features: PageValues and SpecialDays
   b. Filter Method : Quasi Constant, Chi Square, Univariate Selection, and Mutual Information
   c. Embedded Method : Lasso
   d. Before Feature Selection: 51 features, After feature selection: 16 features.

7. **Handling Imbalanced Target**
   SMOTE with default sampling_strategy (1:1)