# Online Shoppers Purchasing Intention

**MAJESTIC**


- Theo Jodanta
- Shanna Sinaga
- Refi Fadholi
- Riswan Setiawan
- Fajar Arief
- Faiz Naida

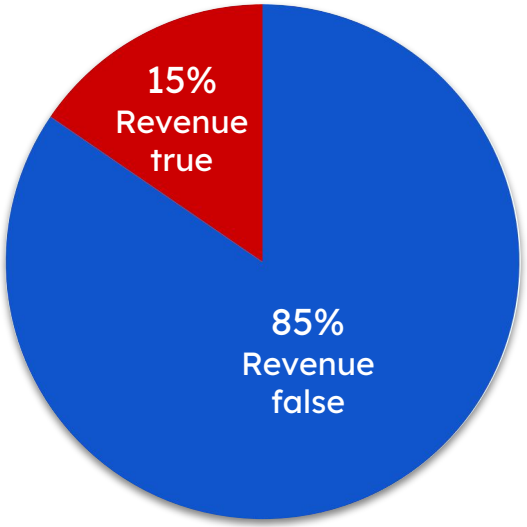

# Table of Content

# 01

# PROBLEM STATEMENT

# WHAT IS THE PROBLEM?

Majestic merupakan suatu perusahaan E-commerce (marketplace) yang menyediakan berbagai macam kebutuhan untuk pelanggan.



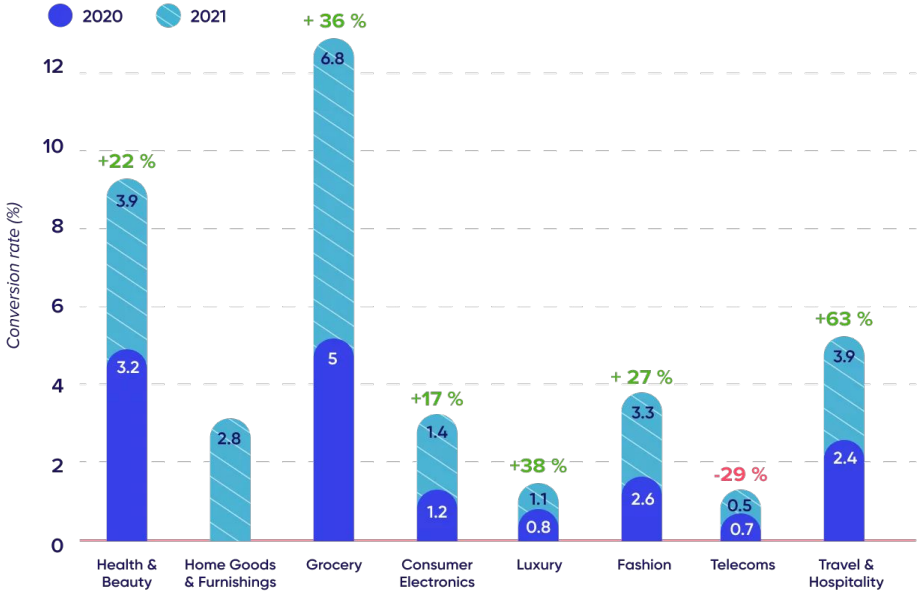

Pada satu tahun terakhir, perusahaan hanya menghasilkan conversion rate (revenue true) sebesar 15% dari pelanggan yang mengunjungi website.

# WHAT IS THE PROBLEM?

Pada masa pandemi (2020-2021) menurut data **Digital Experience Benchmark Report**, conversion rate diberbagai industri e-commerce mengalami peningkatan rata-rata sebesar 28% dikarenakan secara signifikan kebiasaan customers untuk berbelanja beralih ke sistem online.

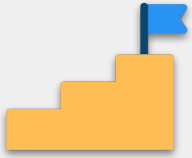
**OPORTUNITY**
**Meningkatkan revenue**

– Average Conversion Rate for eCommerce Industries




**Source** :
Ecommerce Conversion Rate | Avg Conversion Rate - Contentsquare

Kami sebagai Tim Data Scientist akan mencari solusi dalam mengatasi permasalahan Tim Bisnis ini melalui dataset yang tersedia.



## Objectives

- Mendapatkan insight dari pola customer
- Memprediksi pengunjung yang memiliki kecenderungan membeli atau tidak
- Memberikan bisnis rekomendasi yang tepat

## Goals

- Membuat model machine learning yang dapat memprediksi customer yang berpeluang menghasilkan revenue.
- Diharapkan revenue conversion rate dapat meningkat mencapai 28%.

## Matric

Revenue

Conversion

Rate

**Source** :
Ecommerce Conversion Rate | Avg Conversion Rate - Contentsquare

# 02

# EXPLORATORY DATA ANALYSIS

# Tentang Dataset

Dataset memiliki fitur-fitur yang menunjukkan aktivitas website dan aktivitas pembelian pada sebuah online shop.

## Shape
12330 records
18 features

## Missing Values
0

## Duplicates
123

## Data Type
float64, int64, boolean, object

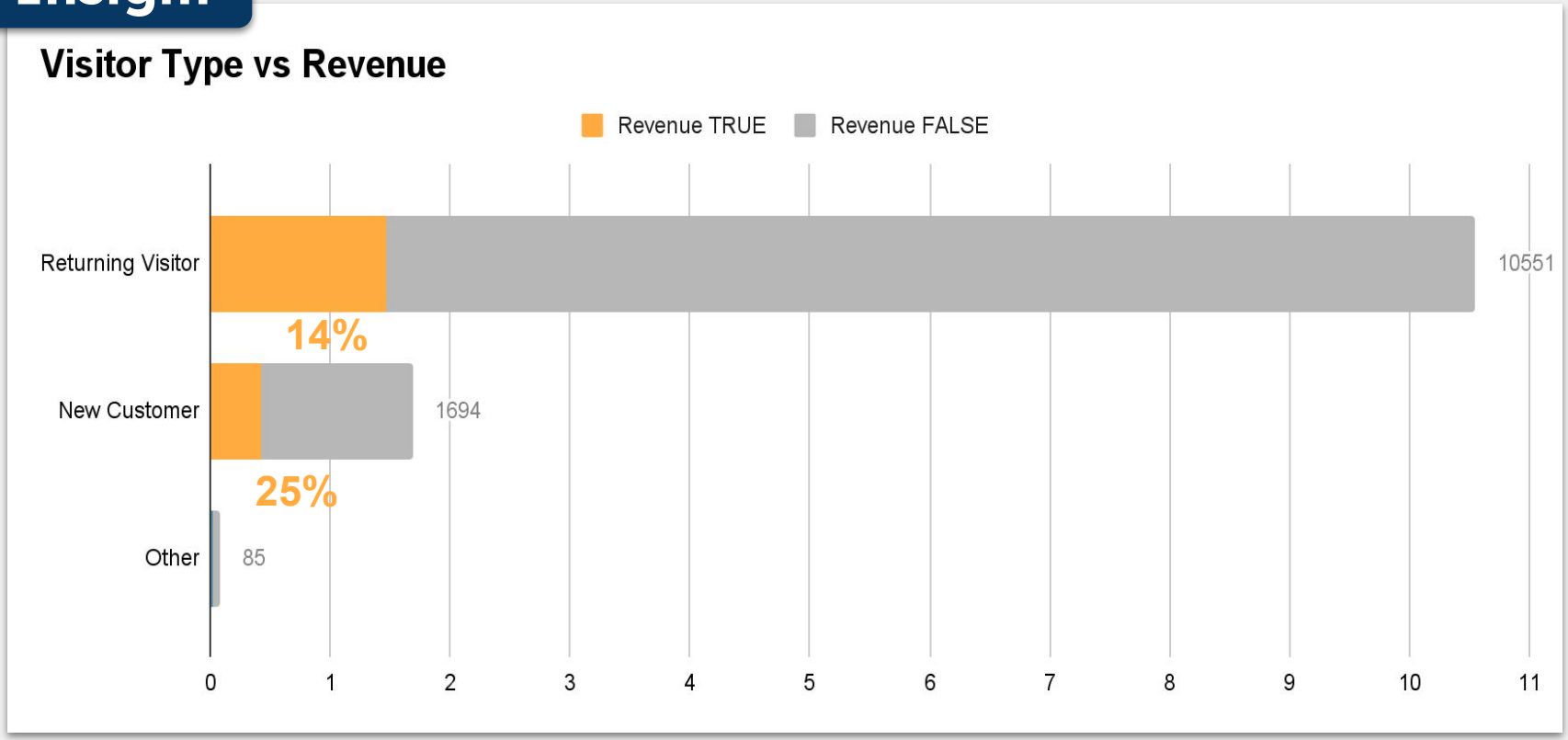
## 10 Fitur Numerikal

Administrative
Administrative Duration
Informational
Informational Duration
ProductRelated
ProductRelated Duration
BounceRates
ExitRates
PageValues
Specialday

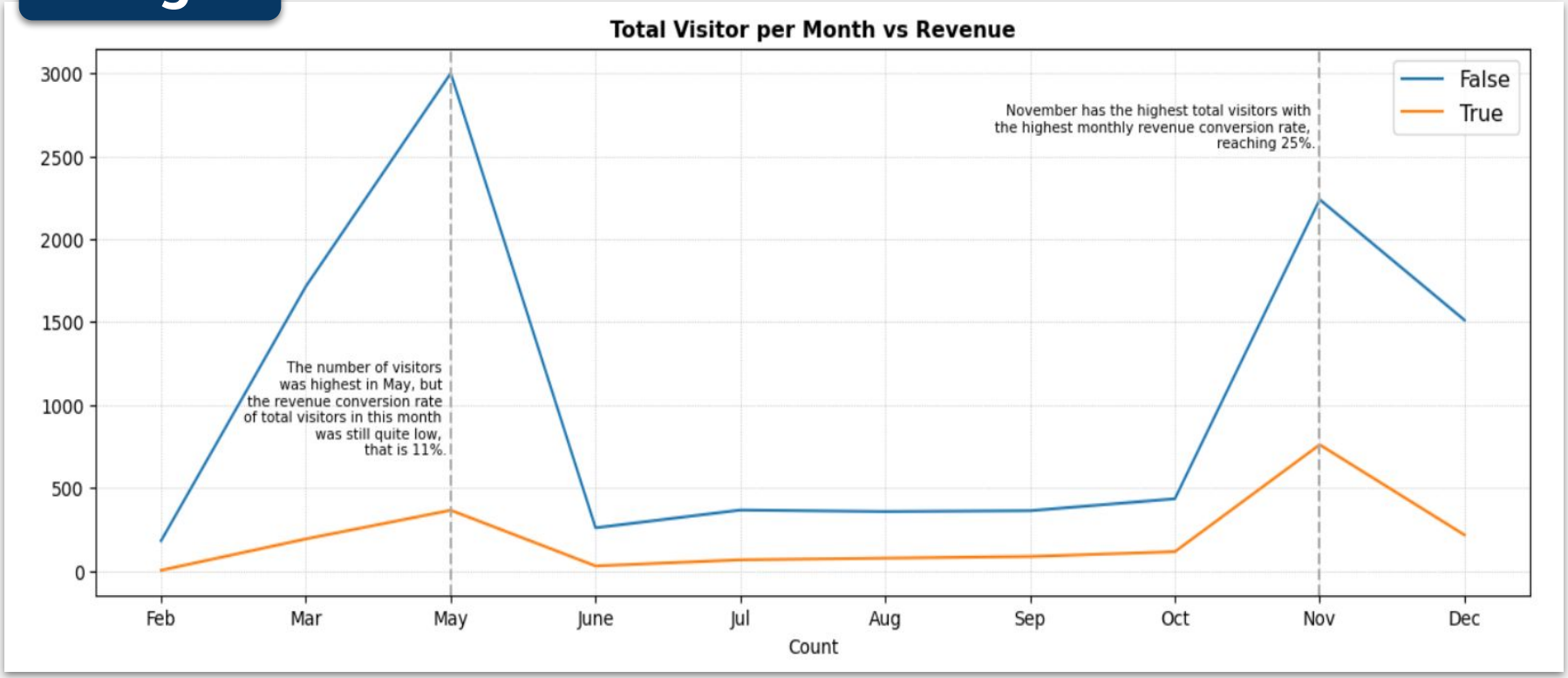
## 8 Fitur Kategorikal

Month
OperatingSystems
Browser
Region
TrafficType
VisitorType
Weekend
**Revenue**

# Insight


Visitor Type vs Revenue
Revenue TRUE
Revenue FALSE
Returning Visitor
10551
14%
New Customer
1694
25%
Other
85
0
1
2
3
4
5
6
7
8
9
10
11

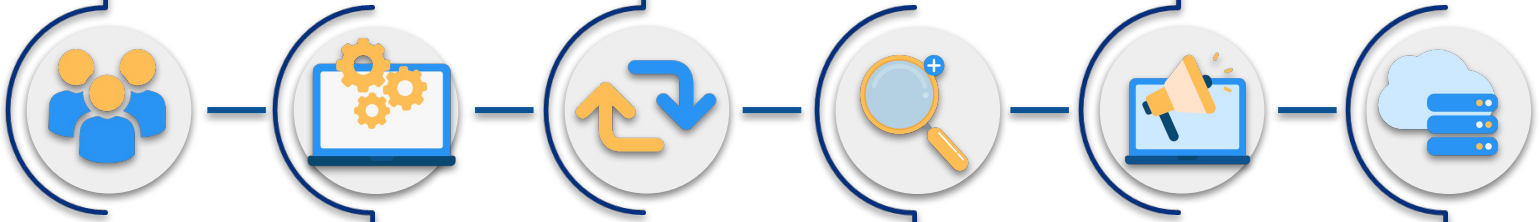
# Insight


Total Visitor per Month vs Revenue
3000
2500
2000
1500
1000
500
0
False
True
November has the highest total visitors with the highest monthly revenue conversion rate, reaching 25%.
The number of visitors was highest in May, but the revenue conversion rate of total visitors in this month was still quite low, that is 11%.
Feb
Mar
May
June
Jul
Aug
Sep
Oct
Nov
Dec
Count

# 03

# DATA PRE-PROCESSING

# Data Pre-Processing



### Handling Nilai, Tipe Data, dan Duplikat
- Drop 125 duplikat record
- Tidak ada NaN

### Handling Outlier
Tidak dilakukan handling outlier

### Feature Transformation
Power Transformer Yeo-johnson

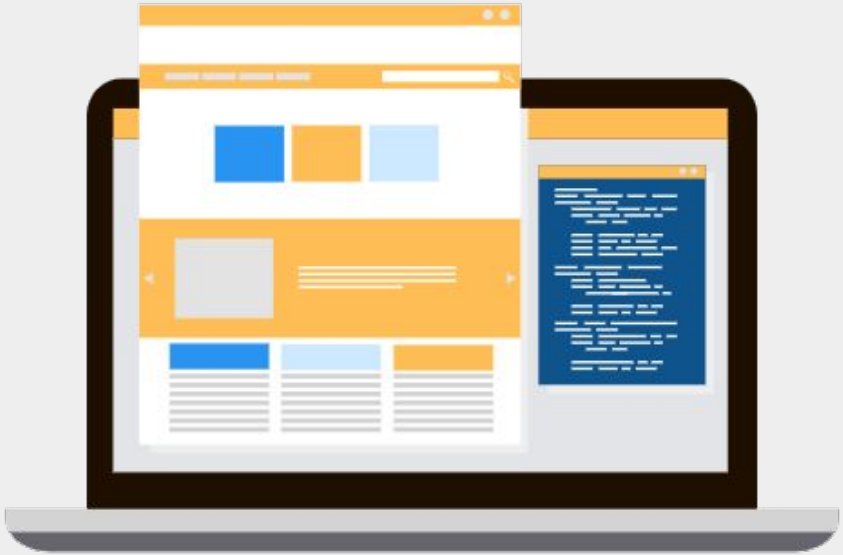
### Feature Encoding
- OHE : Month
- Label : VisitorType, revenue

### Feature Extraction
Membuat fitur baru dari fitur yang ada. Duration per page :
- Administrative
- Informational
- ProductRelated

### Feature Selection
- Drop fitur yang tidak relevan dan redundan

## Feature Selection

- Duration per Page Administrative
- Duration per Page Informational
- ProductRelated
- ExitRates
- PageValues
- Revenue_True
- VisitorType_Returning_Visitor
- SpecialDay



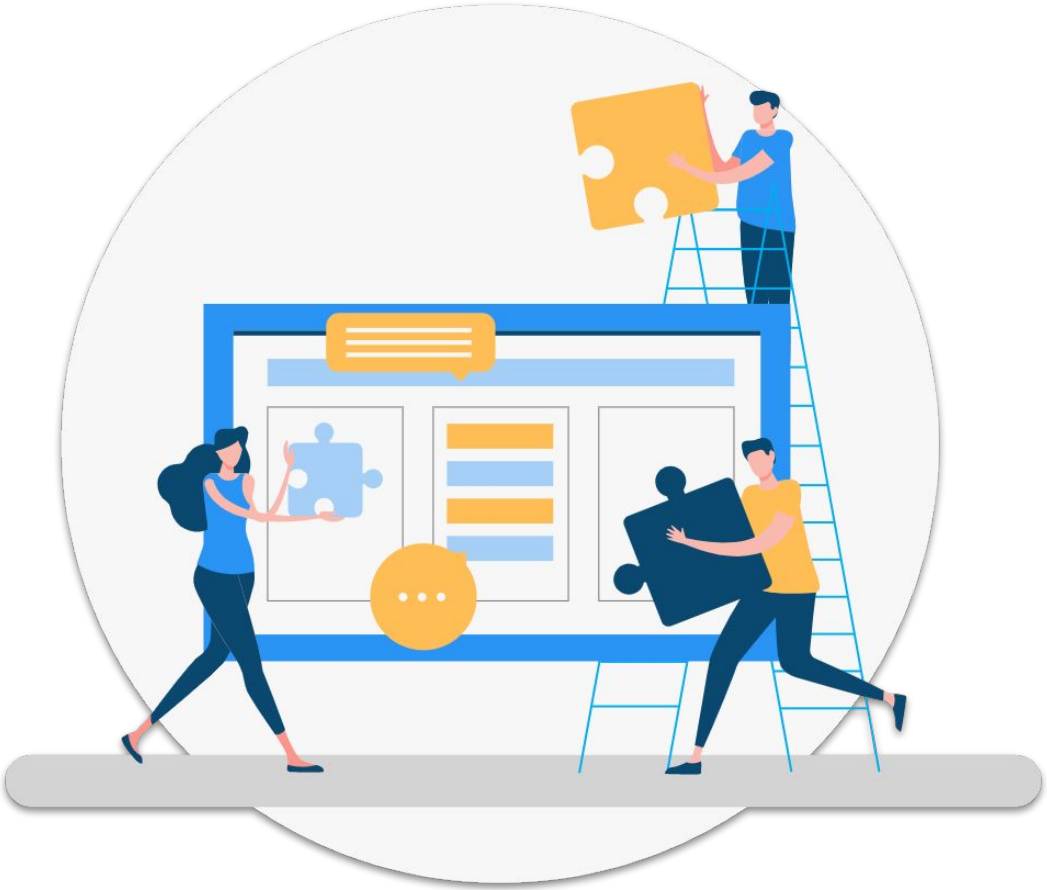
## Split Dataset

**70 % data train**

**30 % data test**

## Class Imbalance

**SMOTE**



# 04

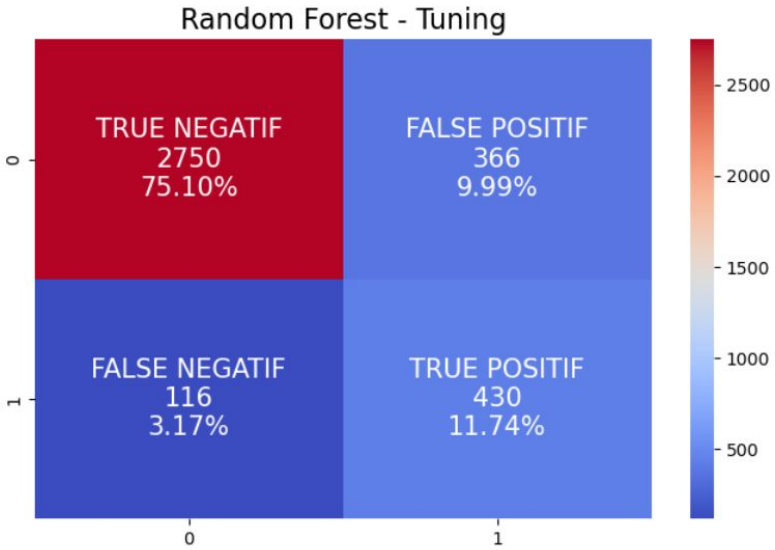
# MODELING

# Classification Model

| | Logistic Regression | KNN | Decision Tree | Random Forest | AdaBoost | XGBoost |
|---|---|---|---|---|---|---|
| Acuraccy | 0.87 | 0.86 | 0.84 | 0.88 | 0.88 | 0.88 |
| Precision | 0.56 | 0.52 | 0.48 | 0.58 | 0.56 | 0.58 |
| Recall | 0.79 | 0.74 | 0.62 | 0.71 | 0.76 | 0.68 |
| F1-Score | 0.65 | 0.61 | 0.54 | 0.64 | 0.65 | 0.63 |
| AUC | 0.89 | 0.86 | 0.74 | 0.90 | 0.89 | 0.90 |

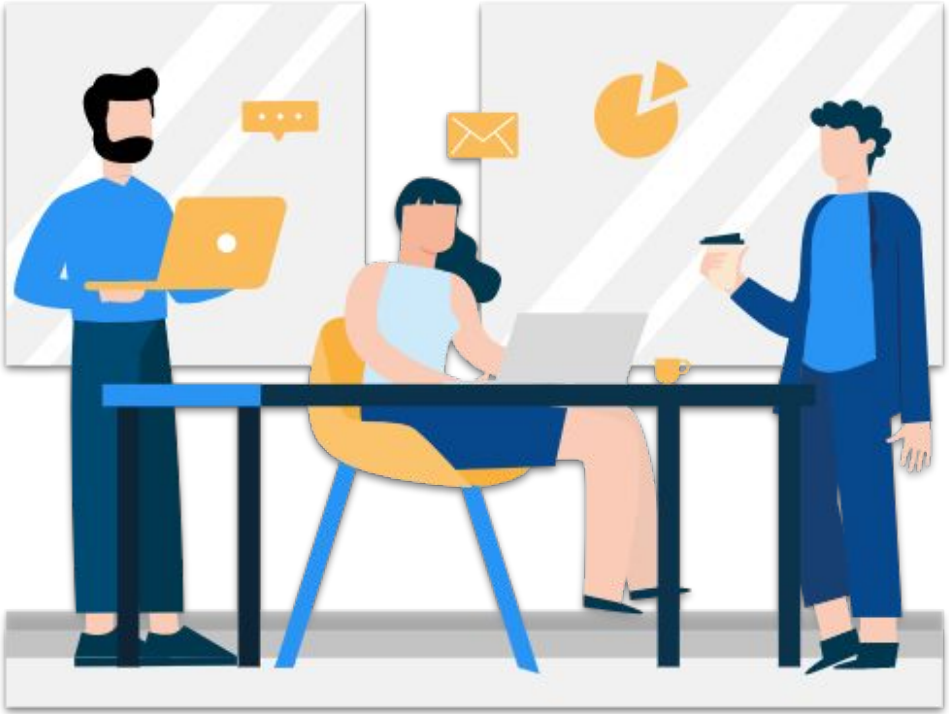
# Final Model - Random Forest (Hyperparameter tuning)

| Train Accuracy | Test Accuracy | Train - Recall | Test - Recall | Train ROC -AUC | Test ROC - AUC |
|---|---|---|---|---|---|
| 0.86 | 0.87 | 0.80 | 0.79 | 0.90 | 0.90 |




Alasan pemilihan model:

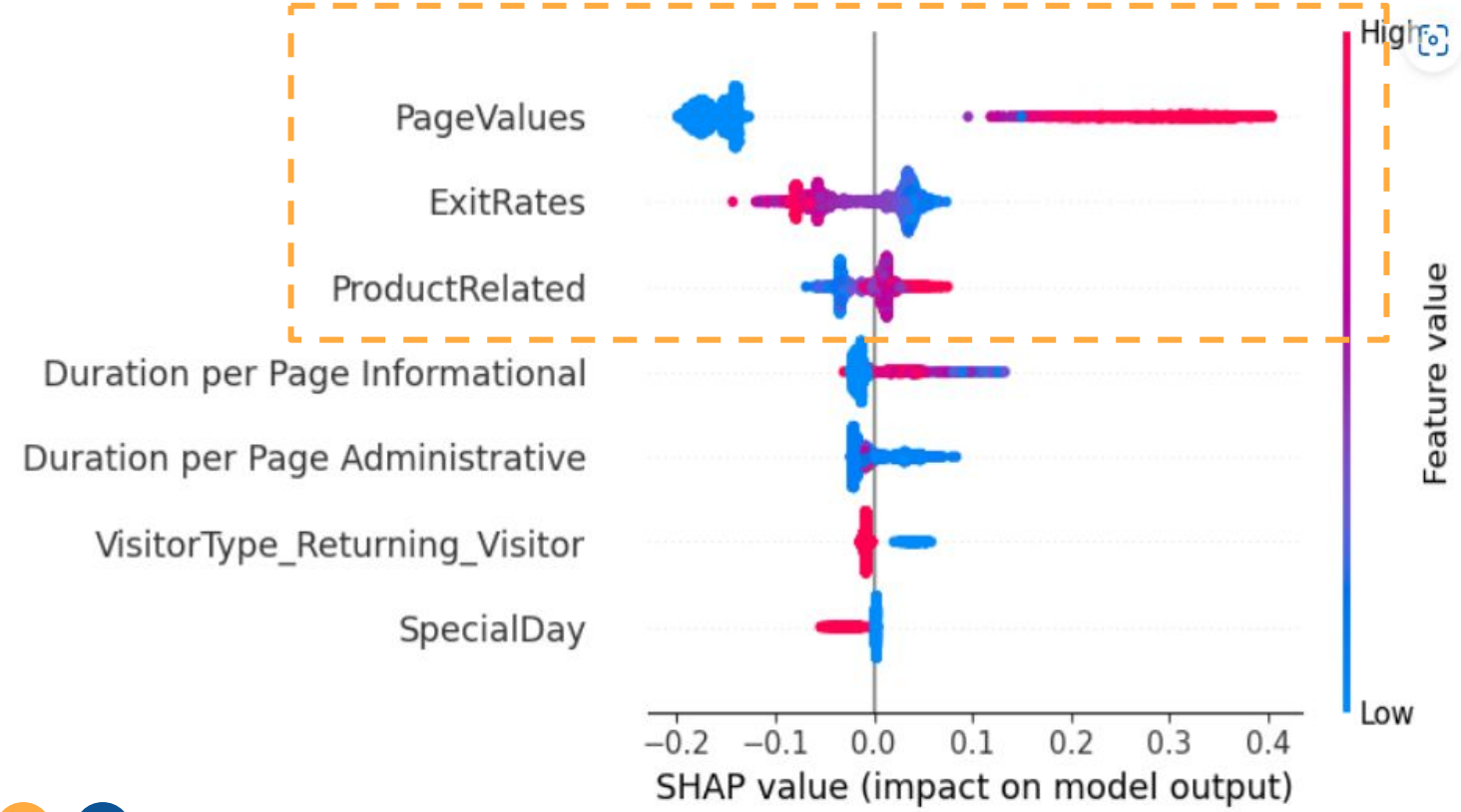
1. Nilai metrik yang tidak terlalu tinggi antara data train dan data test
2. Nilai metrik Accuracy, Recall, ROC-AUC yang cukup tinggi



# 05

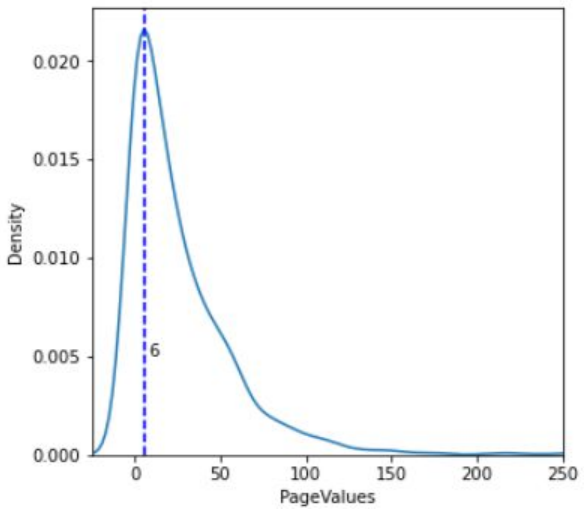
# BUSINESS INSIGHT & RECOMMENDATION

# Feature Importance

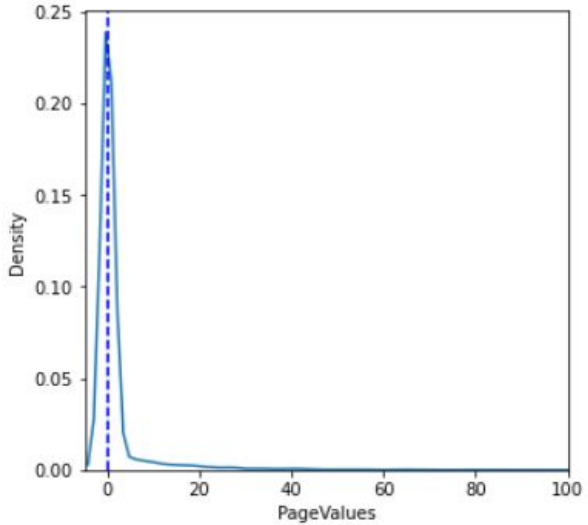


# Business Insight - Page Values

## Data Revenue = True



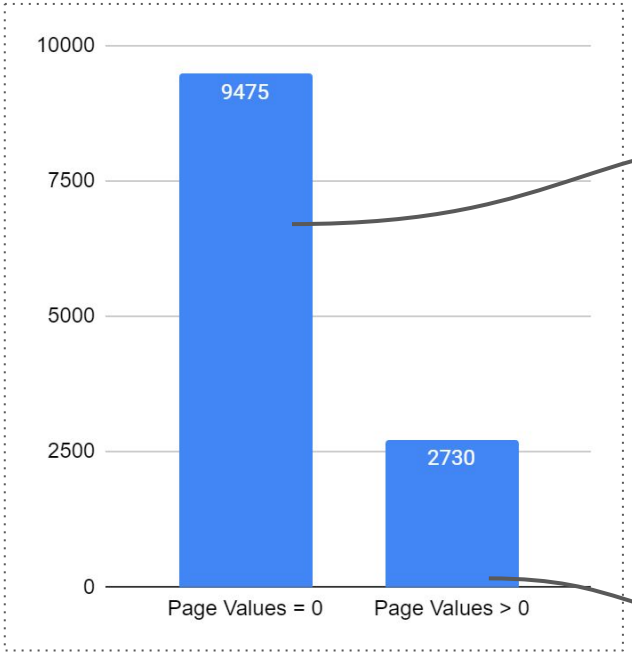

## Data Revenue = False




Visitor yang melakukan purchase (Revenue = True) cenderung mengunjungi halaman dengan nilai Page Values yang lebih besar dibandingkan visitor yang tidak melakukan purchase (Revenue = False)

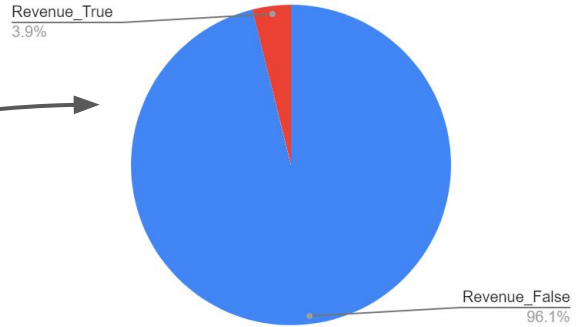
# Business Insight - Page Values

**Grafik Jumlah Page Values = 0 dan Page Values > 0**

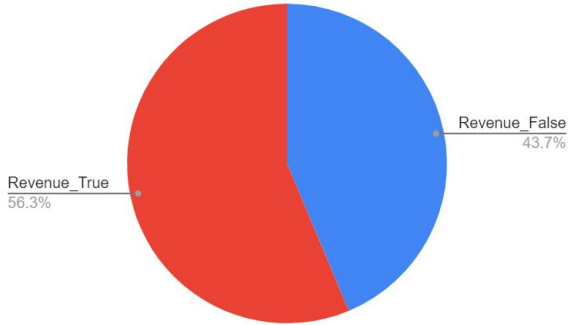



Page Values = 0 lebih banyak (9475 visitor) dibandingkan Page Values > 0 (2730 visitor)

**Persentase Revenue False dan True dengan Page Values = 0**



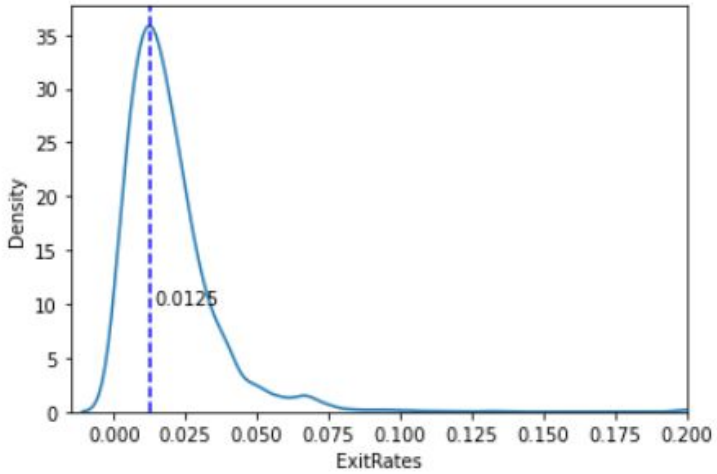

**Persentase Revenue False dan True dengan Page Values > 0**




Page Values > 0 memiliki persentase Revenue True yang lebih besar dibandingkan Page Values = 0

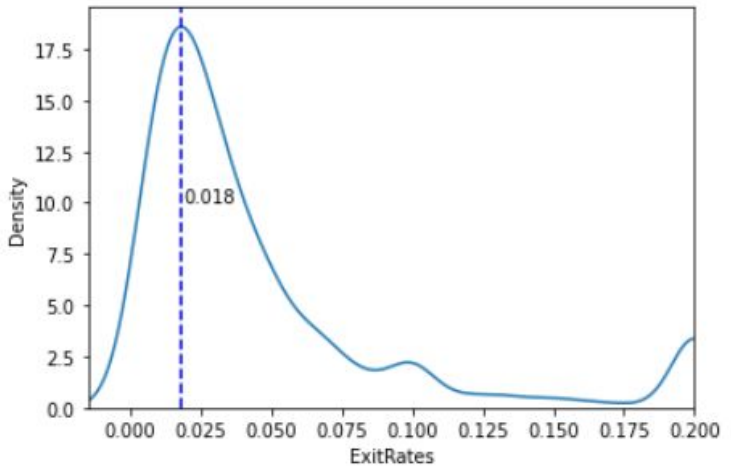
# Business Insight - Exit Rates

## Data Revenue = True




Visitor yang melakukan purchase (Revenue = True) cenderung memiliki Exit Rates yang lebih kecil dibandingkan visitor yang tidak melakukan purchase (Revenue = False)

## Data Revenue = False



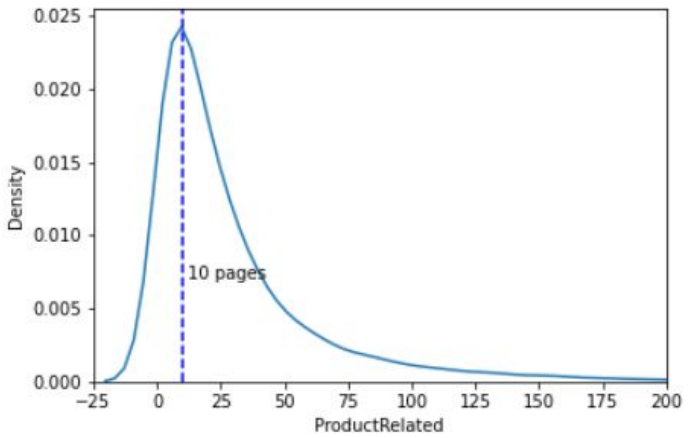

Rata-rata Exit Rates untuk sebuah website e-commerce berada di bawah 25%*, Untuk toko E-Commerce ini, nilai maksimal Exit Rates adalah 20%. Untuk itu, nilai Exit Rates masih dapat diterima.

*https://upqode.com/bounce-rate-vs-exit-rate/

# Business Insight - Product Related

Rata-rata page views untuk website seluruh industri adalah 5.*

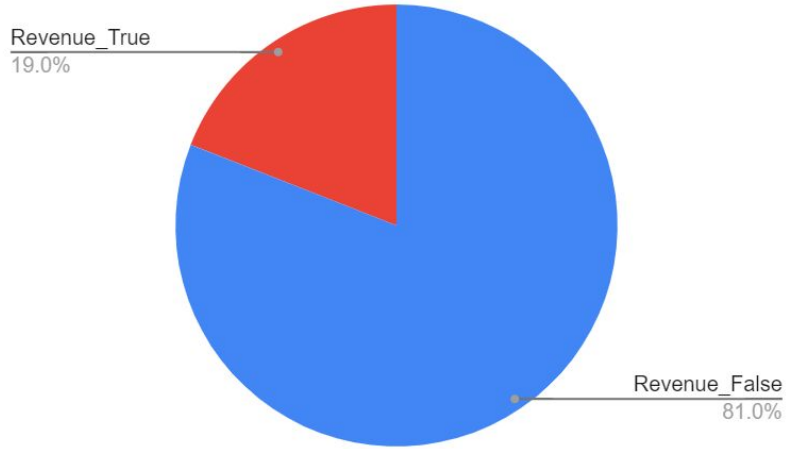
**Grafik Persebaran Laman Product Related**




Berdasarkan data di atas, nilai page views Product Related sebagian besar sudah melebihi dari rata-rata sumber.

Probabilitas seseorang untuk membeli sangat tinggi apabila orang tersebut menghabiskan waktu paling tidak 50 detik pada sebuah halaman produk**

**Grafik Persentase False dan True Revenue pada Data dengan Durasi per Product Related >= 50**

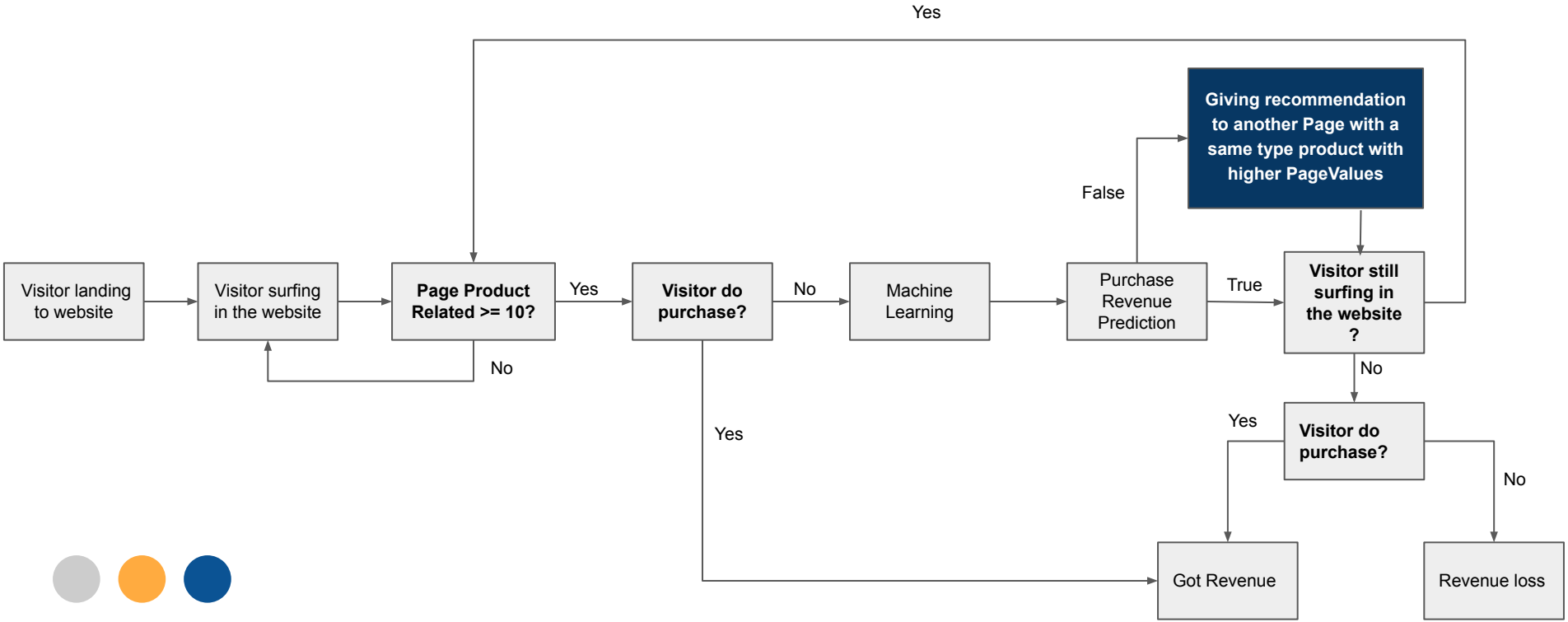



*Pages Per Session | MetricHQ (klipfolio.com)

**'Average Time on Page' & Its Impact on Purchase Probability for E-commerce | by Cappasity | Cappasity Blog | Medium

# Machine Learning Prediction Workflow

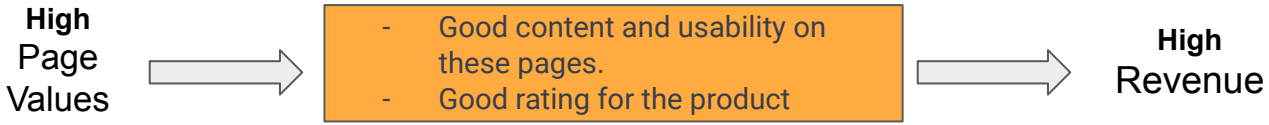
**Early Purchase Prediction (Machine Learning** untuk memprediksi pengunjung untuk membeli atau tidak sebelum pengunjung melakukan pembelian)



# Business Recommendation

- **Business Recommendation for Machine Learning**

**Pemberian notifikasi rekomendasi laman produk yang berjenis sama dengan Page Values yang lebih tinggi**




Kecenderungan visitor untuk melakukan purchase dipengaruhi juga oleh informasi rekomendasi yang berdasarkan oleh penilaian subjektif (seperti online feedback)*

- **Business Recommendation from Insight**

1. Pemberian diskon dan pengadaan event pada bulan-bulan tertentu (Mei, Maret, November, dan Desember
2. Dikarenakan persentase revenue  New_customer lebih besar (25%) dibandingkan Returning_Visitor (14%), maka direkomendasikan untuk meningkatkan traffic untuk New_customer,
   dengan cara:
   1. Pemberian diskon khusus pengguna baru
   2. Melakukan iklan di beberapa media untuk meningkatkan ketahuan masyarakat terhadap website E-Commerce

*Y. -C. Ku and Y. -M. Tai, "What Happens When Recommendation System Meets Reputation System? The Impact of Recommendation Information on Purchase Intention," *2013 46th Hawaii International Conference on System Sciences*, 2013, pp. 1376-1383, doi: 10.1109/HICSS.2013.605.

# Simulasi

| Before ML | | After ML |
|---|---|---|
| Rp 57.240.000 | Revenue | Rp 123.337.628 |
| 15% | Conversion Rate | 34% |
| 1908 | Visitor Revenue True | 4111 |

**Potential Revenue** **53,59%**

## Asumsi

- Recommendation implementation success rate : **25%**
- Commision revenue/transaction : **Rp 30.000**
- Every visitor just **purchase one product**

# Thank You

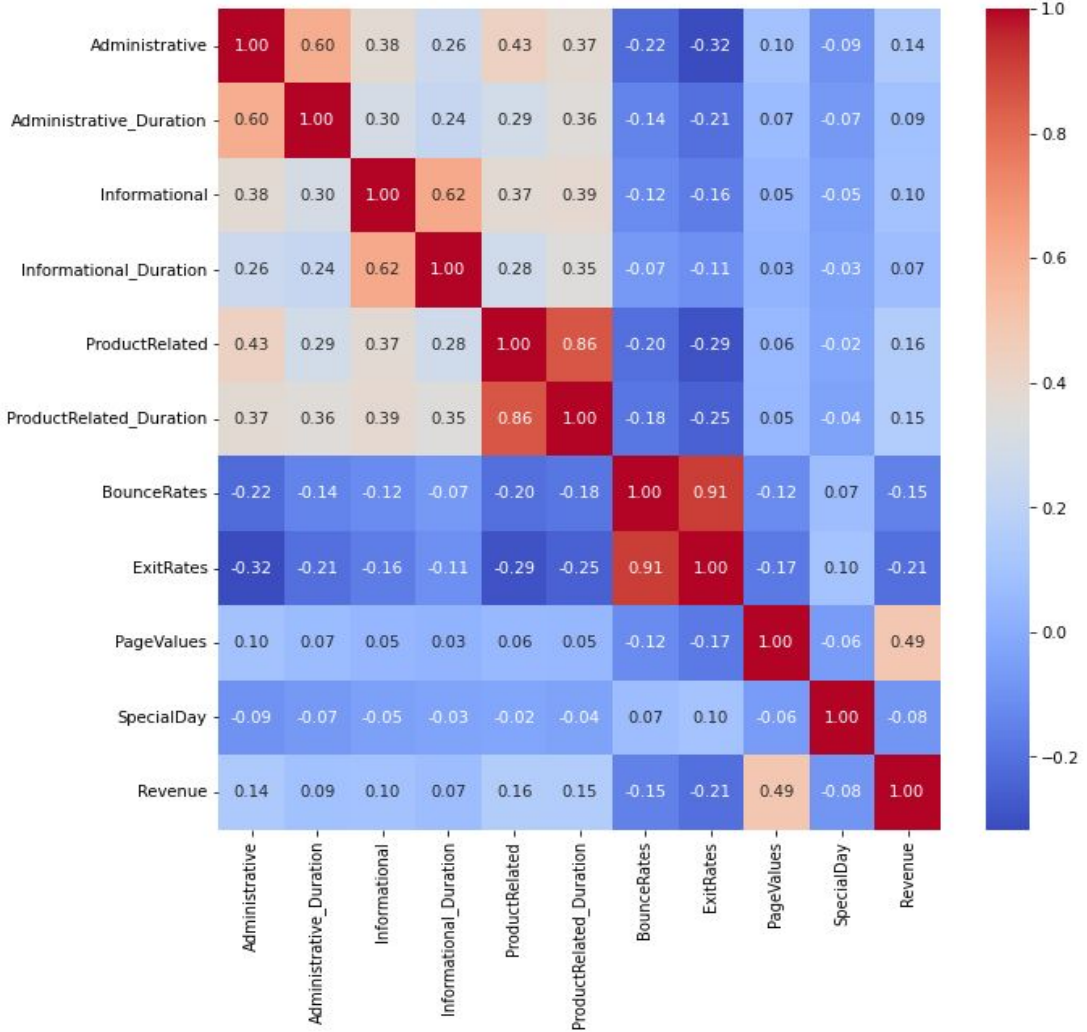
# APPENDIX

# SIMULASI MODEL MECHINE LEARNING

| | A | B | C | D | E |
|---|---|---|---|---|---|
| | Sebelum ML | | | | |
| | Revenue/transaksi | Rp30,000 | | | |
| | Jumlah visitor | 12205 | *(tanpa duplikat)* | | |
| | Jumlah visitor Yes | 1908 | | | |
| | Conversion rate | 15.63% | | | |
| | **Total revenue** | **Rp57,240,000** | | | |
| | | | | | |
| | Sesudah ML | | | | |
| | Revenue/transaksi | Rp30,000 | | | |
| | Jumlah visitor | 12205 | | | 18.05% |
| | Promotion effective rate | 25% | *(ada referensi)* | | |
| | **True Negative prediction** | 75.10% | | | |
| | **True Positive prediction** | 11.74% | | | |
| | **False Positive** | 9.99% | | **Peningkatan revenue** | **53.59%** |
| | **False Negative** | 3.17% | | | |
| | Jumlah visitor Yes | 4111.25425 | | | |
| | Conversion rate | 33.69% | | | |
| | **Total revenue** | **Rp123,337,628** | | | |

# Analisis Multivariate

1. PageValues memiliki korelasi yang tinggi terhadap fitur target yaitu Revenue. Semakin tinggi nilai PageValue, maka semakin tinggi juga kemungkinan pelanggan untuk membeli.
2. Sedangkan BounceRates dan ExitRates memiliki nilai korelasi negatif terhadap Revenue, artinya semakin kecil nilai kedua fitur tersebut maka revenue akan semakin tinggi.
3. Beberapa fitur yang memiliki multikorenialitas diantaranya adalah :
    - ProductRelated dengan ProductRelated_Duration
    - Adminisitrative dengan Adminisitrative_Duration
    - Informational dengan Informational_Duration
    - BounceRates dengan ExitRate

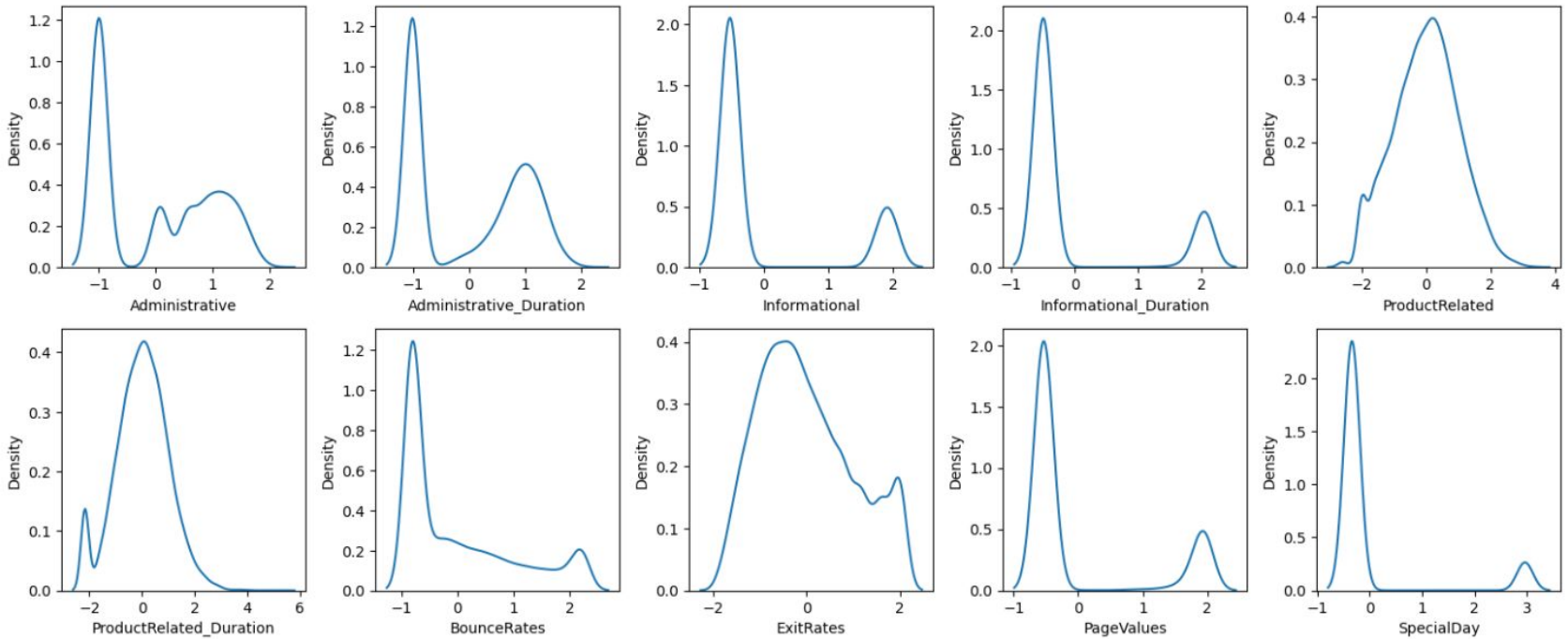


# Transformasi Fitur

```
# transformasi data

from sklearn.preprocessing import PowerTransformer

for x in nums:
    pt = PowerTransformer(method='yeo-johnson')
    df[x] = pt.fit_transform(df[x].to_frame())
```



## E ) Feature Extraction

In [29]:
```
# memilih feature dengn korelasi tinggi dengan Revenue

x = corrmat['Revenue_True']
result = x[(x>0.05)|(x<-0.05)] # korelasi lebih dari 0.5
result
```

Out[29]:
```
Administrative                  0.164376
Administrative_Duration         0.164306
Informational                   0.110966
Informational_Duration          0.107878
ProductRelated                  0.196981
ProductRelated_Duration         0.211123
BounceRates                    -0.172585
ExitRates                      -0.249863
PageValues                      0.611599
SpecialDay                     -0.088071
VisitorType_Returning_Visitor  -0.102694
Revenue_True                    1.000000
Name: Revenue_True, dtype: float64
```

In [30]:
```
df['Duration per Page Administrative'] = df['Administrative_Duration'] / df['Administrative']
df['Duration per Page Informational'] = df['Informational_Duration'] / df['Informational']
df['Duration per Page ProductRelated'] = df['ProductRelated_Duration'] / df['ProductRelated']
```

# 2. Split Dataset

In [9]:
```
# Split terlebih dahulu sebelum di oversampling
from sklearn.model_selection import train_test_split
X_train, X_test,y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
```

In [10]:
```
# Data Train Didapatkan oversampling dengan SMOTE
from imblearn import under_sampling, over_sampling
X_train, y_train = over_sampling.SMOTE(0.5).fit_resample(X_train, y_train)
print('SMOTE')
print(pd.Series(y_train).value_counts())
```

```
SMOTE
0    7181
1    3590
Name: Revenue_True, dtype: int64
```

## Hyperparamete Tuning Random Forest

In [27]:

```
# tuning hyperparameter RF + oversampling
from sklearn.model_selection import RandomizedSearchCV, GridSearchCV

n_estimators = [int(x) for x in np.linspace(125, 200, 15)]
criterion = ['gini', 'entropy']
max_depth = [int(x) for x in np.arange(3, 5)]
min_samples_split = [int(x) for x in np.linspace(1000, 1200 ,20)]
min_samples_leaf = [int(x) for x in np.linspace(200, 300, 20)] # min_samples_leaf
hyperparameters = dict(n_estimators=n_estimators, criterion=criterion, max_depth=max_depth,
                       min_samples_split=min_samples_split, min_samples_leaf=min_samples_leaf)

rf = RandomForestClassifier(random_state=42)
rf_tuned = RandomizedSearchCV(rf, hyperparameters, cv=5, scoring = 'roc_auc')
rf_tuned.fit(X_train, y_train)
eval_classification(rf_tuned)
```

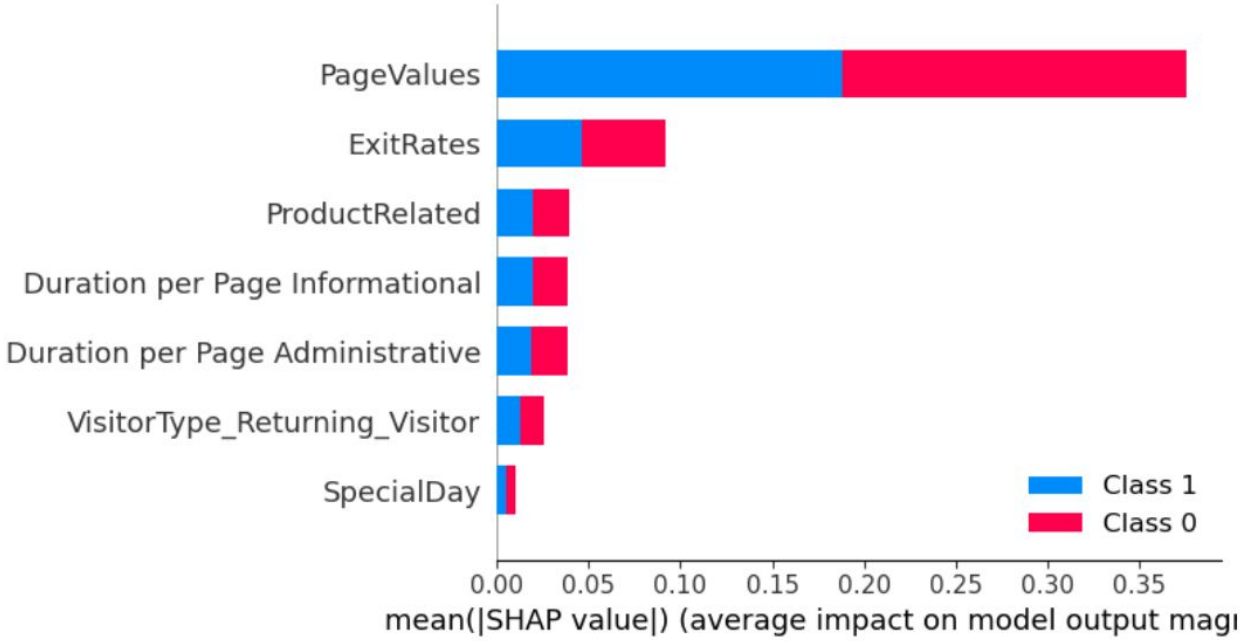
```
Accuracy (Test Set): 0.87
Accuracy (Train Set): 0.86
Precision (Test Set): 0.54
Precision (Train Set): 0.78
Recall (Test Set): 0.79
Recall (Train Set): 0.80
F1-Score (Test Set): 0.64
F1-Score (Train Set): 0.79
roc_auc (test-proba): 0.90
roc_auc (train-proba): 0.91
roc_auc (crossval test): 0.904899775257283
roc_auc (crossval train): 0.9082363056993582
```

[59]:
```
import shap

model = RandomForestClassifier(n_estimators=125, max_depth=4)
model.fit(X_train, y_train)

explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
```

[60]:
```
# bar
shap.summary_plot(shap_values, features=X_test, feature_names=X_test.columns, plot_type='bar')
```



```
# beeswarm
shap.summary_plot(shap_values[1], X_test) # Summary shap value terhadap label positive
```