



Team Seaborn

OPTIMIZING CUSTOMER CONVERSION RATES IN E-COMMERCE

A Data-Driven Analysis of Customer Behavior and Factors Influencing Conversion



Mentor

Dino Febriyanto



Fajar



Wafi



Donni



Fildzah



Rofik



Bintang



Widya



Roya

Content

01

Problem

02

Goals

03

Framework & Method

04

EDA

05

Data Preprocessing

06

Modelling

07

Business Impact

08

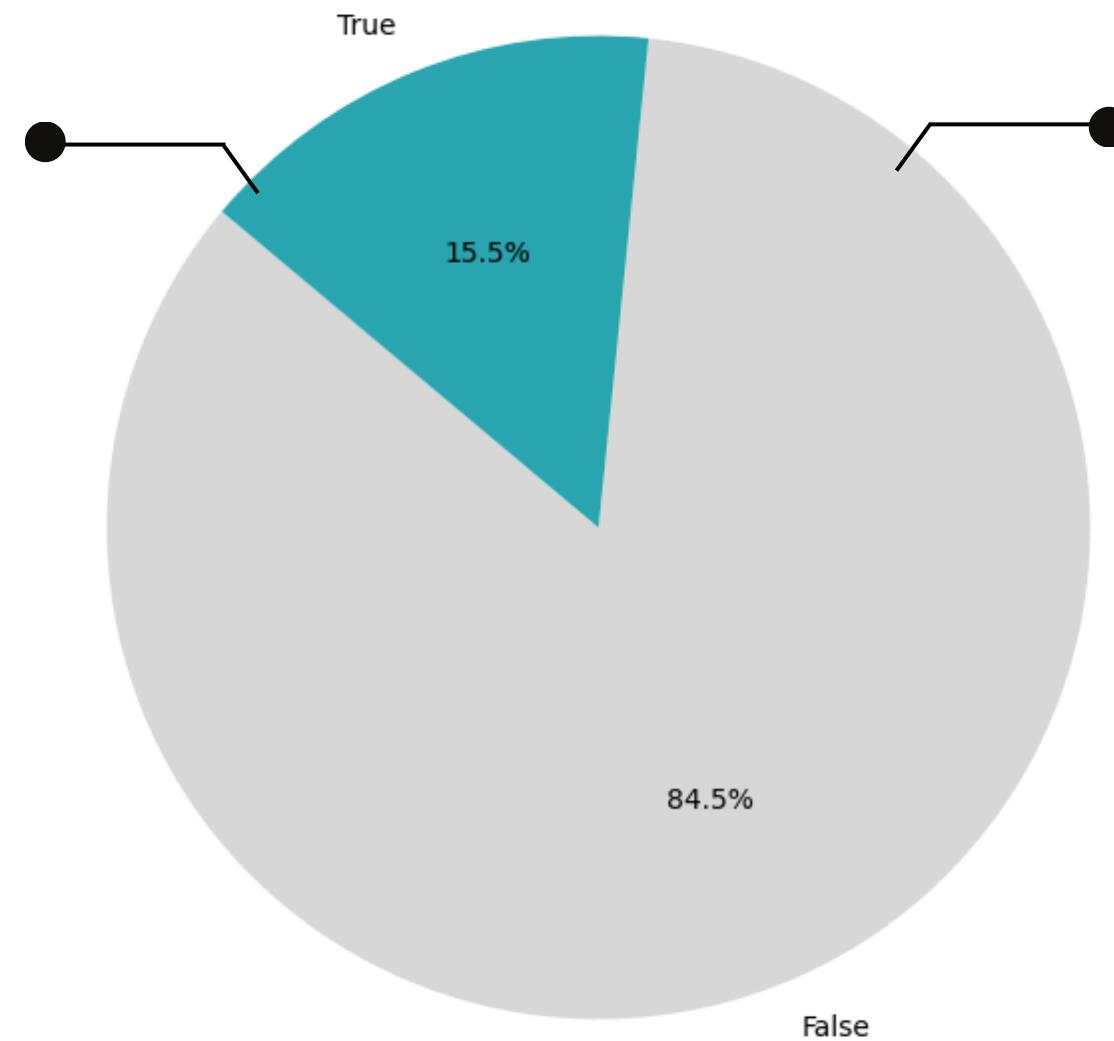
Insight &
Recommendation

Problem

ShopEase adalah platform e-commerce terkemuka yang menyediakan berbagai produk kepada pelanggan.

Background

15.5% pelanggan berhasil terkonversi, dari seluruh orang yang berkunjung di website perusahaan. Dengan rata-rata konversi pelanggan sebesar 14.45%.



Problem

Mempertahankan dan meningkatkan angka konversi tetap menjadi tantangan di tengah persaingan ketat dan perubahan perilaku konsumen.

Sayangnya, perusahaan masih belum memiliki pemahaman mendalam tentang faktor-faktor yang mempengaruhi konversi pelanggan.

Urgensi

Tanpa pemahaman faktor konversi dan tool prediktif, perusahaan beresiko kehilangan peluang untuk mengoptimalkan strategi pemasaran dan menghadapi potensi penurunan konversi pelanggan.

Goal

01

Mengidentifikasi
faktor yang
mempengaruhi
konversi pelanggan

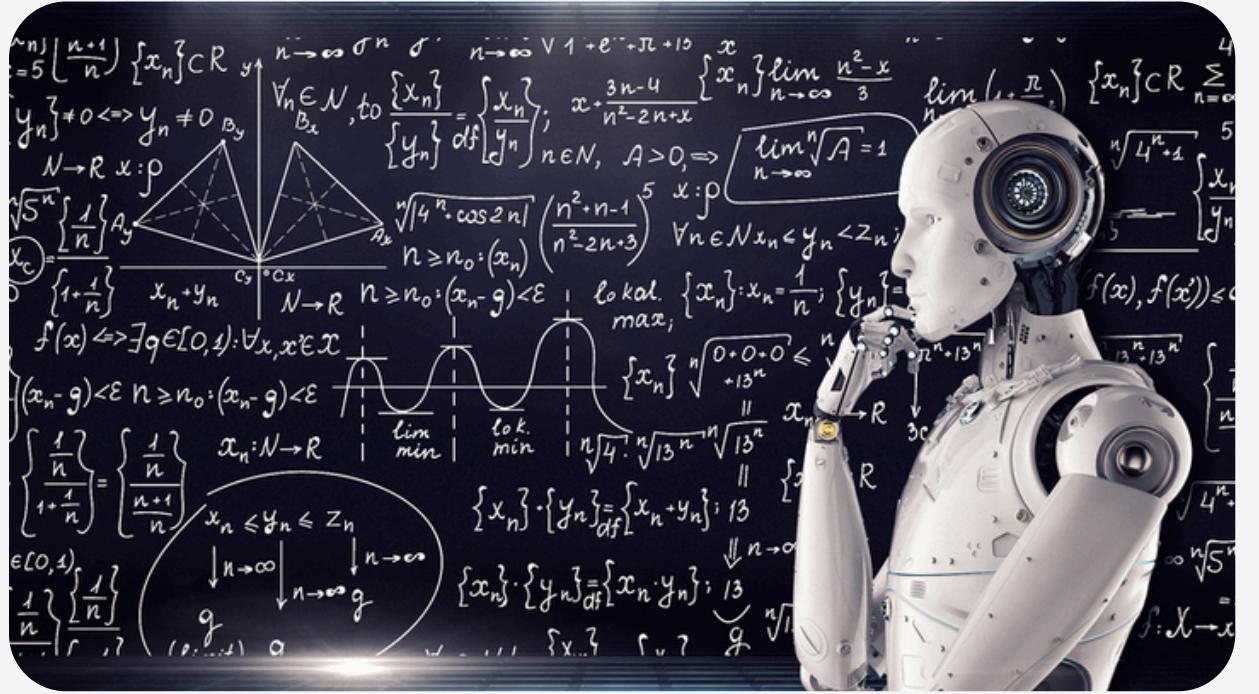
02

Meningkatkan
penjualan
perusahaan

03

Mengantisipasi
penurunan penjualan
lebih lanjut

Method



Machine Learning (ML)

ML adalah cabang dari AI yang memungkinkan sistem dapat belajar dari data, mengidentifikasi pola dan membuat keputusan dengan minim campur tangan manusia. Sistem dapat belajar dengan sedikit atau tanpa pengawasan.

Why Machine Learning (ML) is so popular in the 21st Century

Keunggulan ML

- 61% pengambilan keputusan dalam bisnis menekankan penerapan AutoML
- ML merevolusi sektor pemasaran, dimana mampu meningkatkan dan memperbaiki kepuasan pelanggan hingga lebih dari 10%.
- Menurut McKinsey, perusahaan yang mengadopsi AL dan ML mencatat peningkatan produktifitas hingga 40%
- Boston Consulting Group melaporkan bahwa perusahaan yang menggunakan AI dapat mengurangi biaya operasional hingga 20-30% .

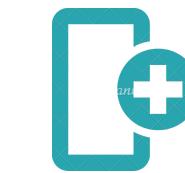
About Dataset

E-Commerce Dataset

Sumber: Sakar, C.O., Polat, S.O., Katircioglu, M.
et al. Neural Comput & Applic (2018).



Terdiri dari 12.946 Baris



Mempunyai 18 kolom, 10 kolom numerik dan 8
kolom kategorik.

Terdiri dari kolom:

1. Administrative
2. Administrative Duration
3. Informational
4. Informational Duration
5. Product Related
6. Product Related Duration
7. Bounce Rates
8. Exit Rates
9. Page Value
10. Special Day
11. Month
12. Operating Systems
13. Browser
14. Region
15. Traffic Type
16. Visitor Type
17. Weekend
18. Revenue (label)

Business Question



- Bagaimana tren konversi Perusahaan?
- Apa saja faktor yang mempengaruhi pengunjung menghasilkan revenue (terkonversi)?
- Bagaimana ML mampu meningkatkan conversion rate perusahaan?
- Apa saja yang dapat dilakukan untuk meningkatkan conversion rate perusahaan?

Business Metric



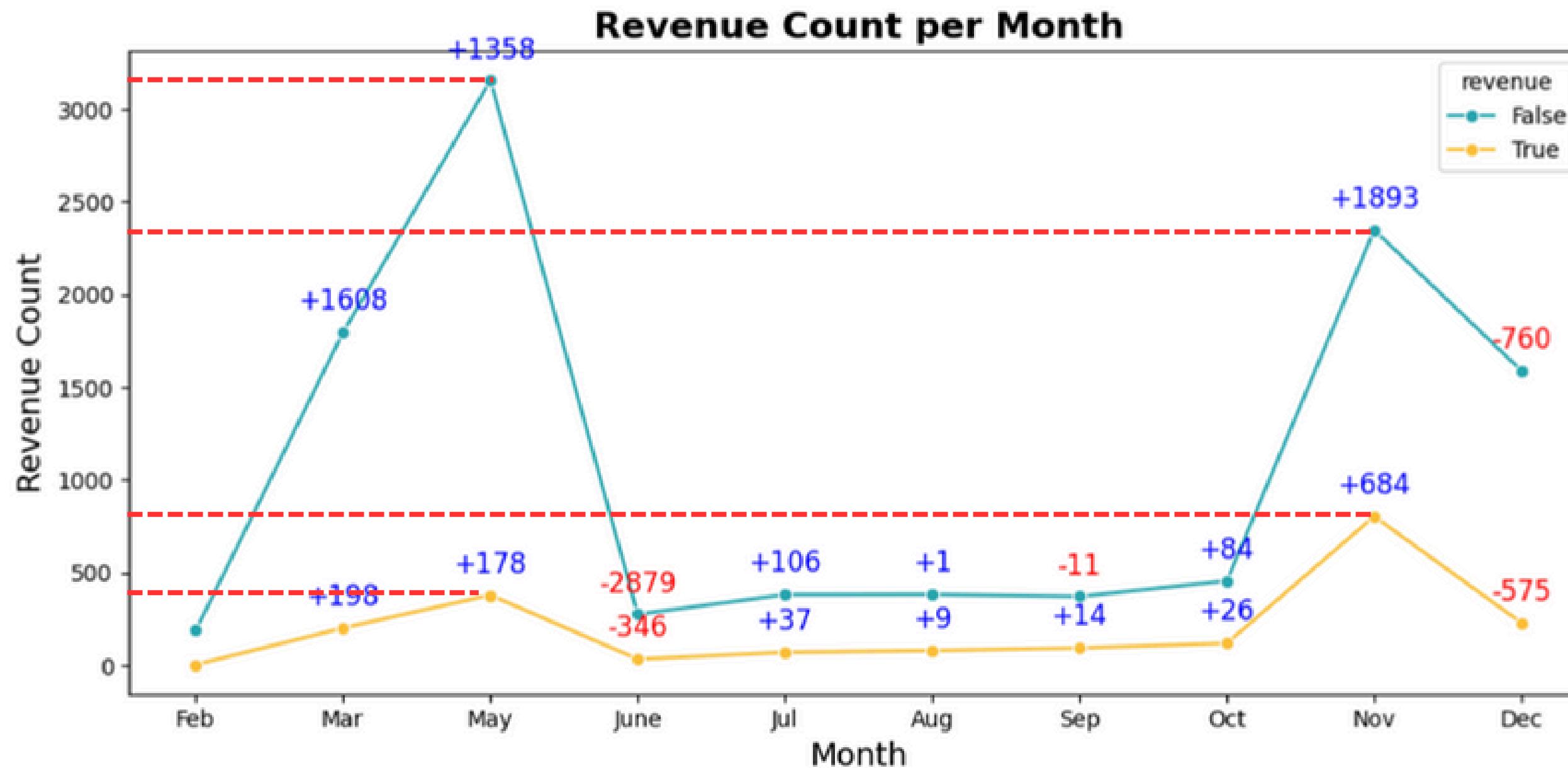
Business metric yang digunakan adalah **Conversion Rate** (Presentase pengunjung yang menghasilkan revenue dari seluruh pengunjung yang ada dalam platform perusahaan).

Explanatory Data Analysis



BI-VARIATE ANALYSIS

PENINGKATAN JUMLAH VISITOR PADA BULAN MEI DAN NOVEMBER



Revenue Count :

May

True : +178 visitor

False : +1.358 visitor

November

True : +684 visitor

False : +1.893 visitor

Perusahaan masih kesulitan dalam mengoptimalkan potensinya (visitor yang tinggi) dan masih memiliki pemahaman yang kurang terkait faktor konversi.

BI-VARIATE ANALYSIS

REVENUE COUNT BY VISITOR TYPE PER MONTH

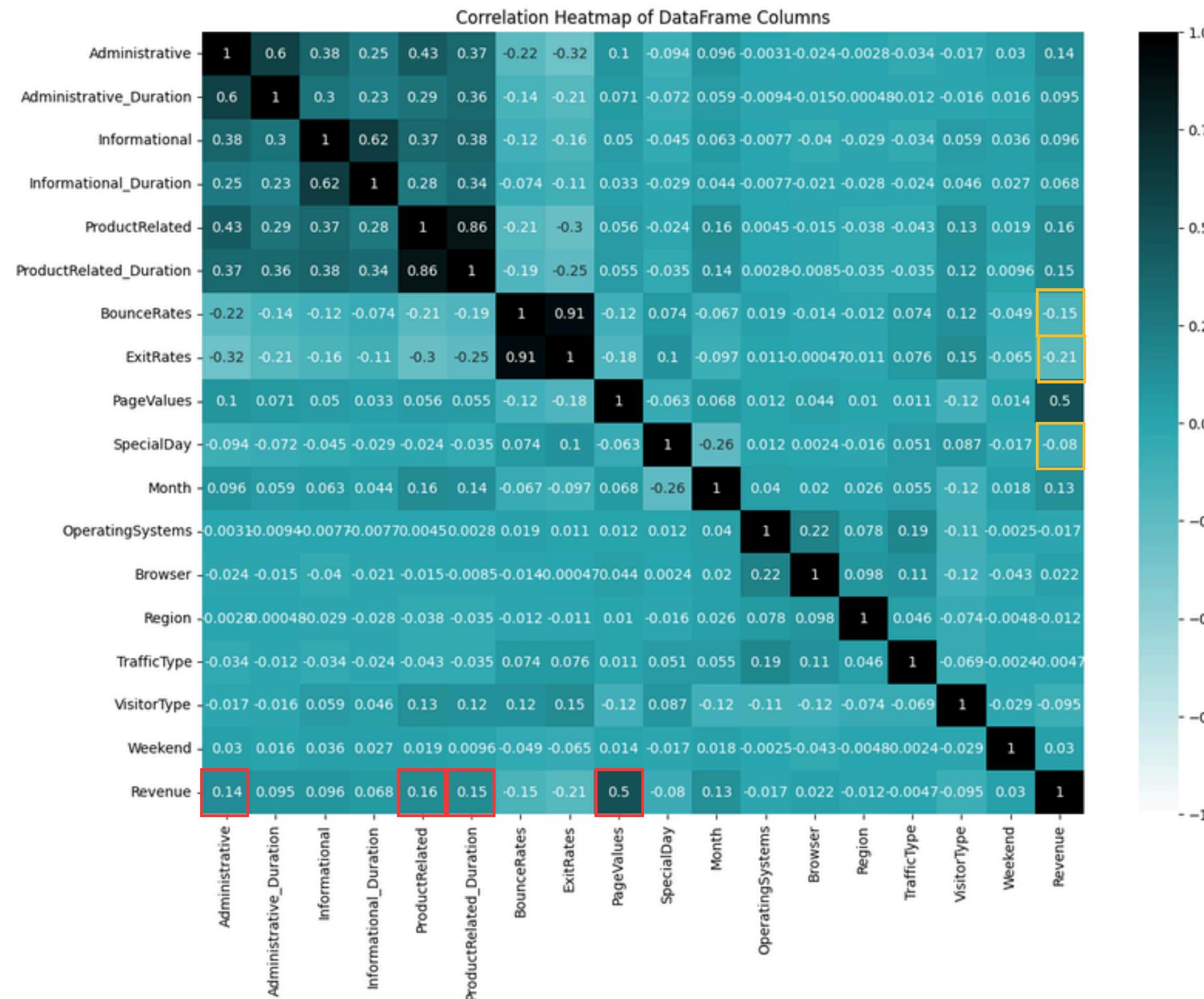


Returning_visitor lebih mendominasi dibandingkan yang lainnya,
namun **GAP New_visitor cukup baik**

Direkomendasikan untuk fokus memperbaiki sistem dan metodenya agar bisa mengkonversi jenis returning visitor

BI-VARIATE ANALYSIS

CORRELATION HEATMAP OF DATAFRAME COLUMNS



Correlation Revenue antar fitur :

Positif (semakin besar nilai = cenderung terkonversi)

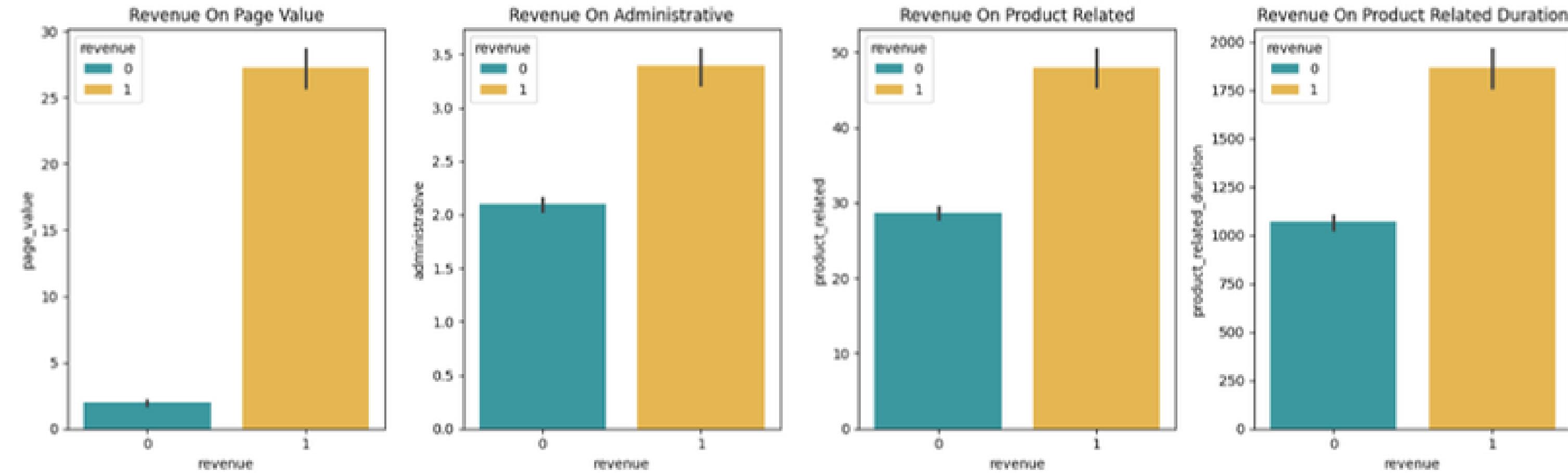
- Pagevalues : 0,5
- ProductRelated : 0,16
- ProductRelated_Duration : 0,15
- Administrative : 0,14

Negatif (semakin rendah nilai = cenderung terkonversi)

- BounceRates : -0,15
- ExitRates : -0,21
- SpecialDay : -0,08

BI-VARIATE ANALYSIS

REVENUE BY CORRELATION FEATURE POSITIVE



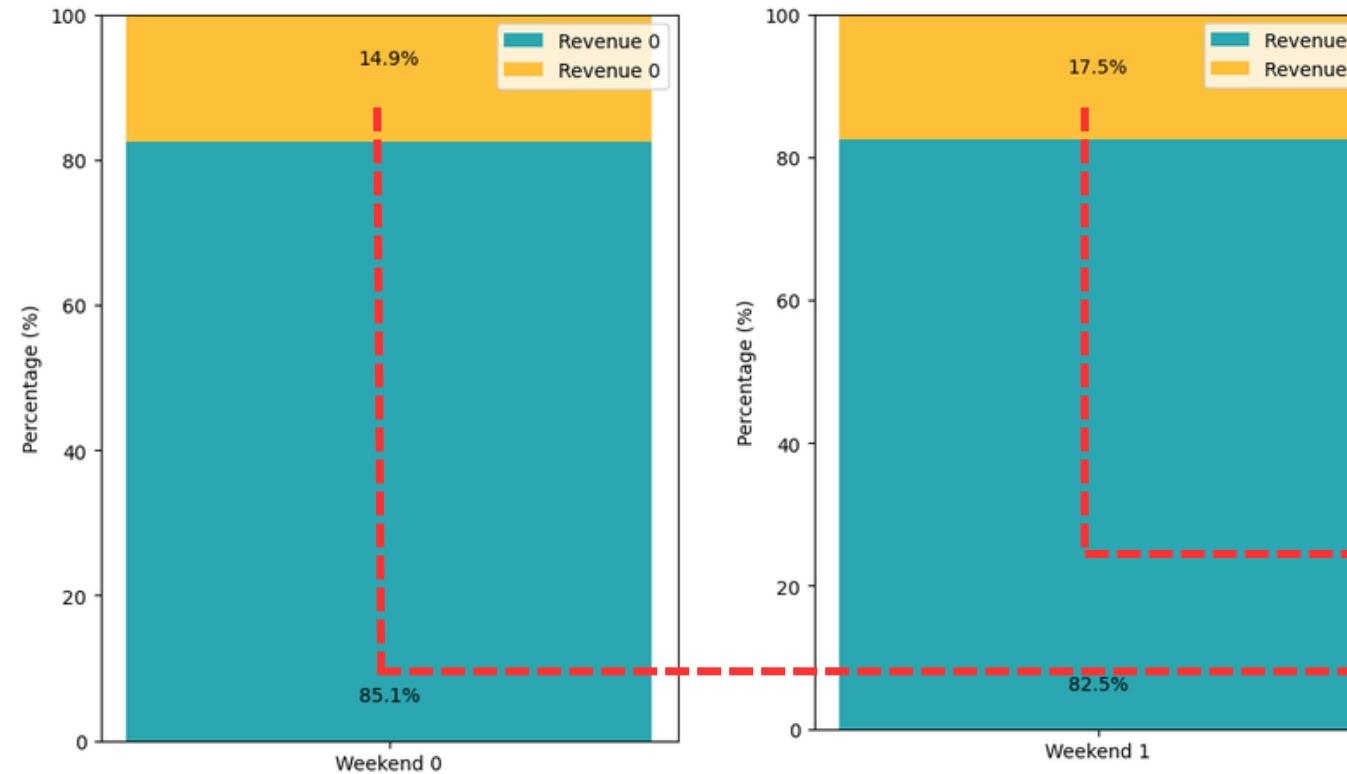
Page Value, Administrative, Product Related dan Product Related Duration
cenderung lebih mudah mengkonversi pelanggan

pelanggan terkonversi = menghasilkan revenue

Fokus pada konten relevan maka dapat mengubah pengunjung menjadi pelanggan setia

BI-VARIATE ANALYSIS

REVENUE BY WEEKDAYS, WEEKEND & SPECIAL DAY



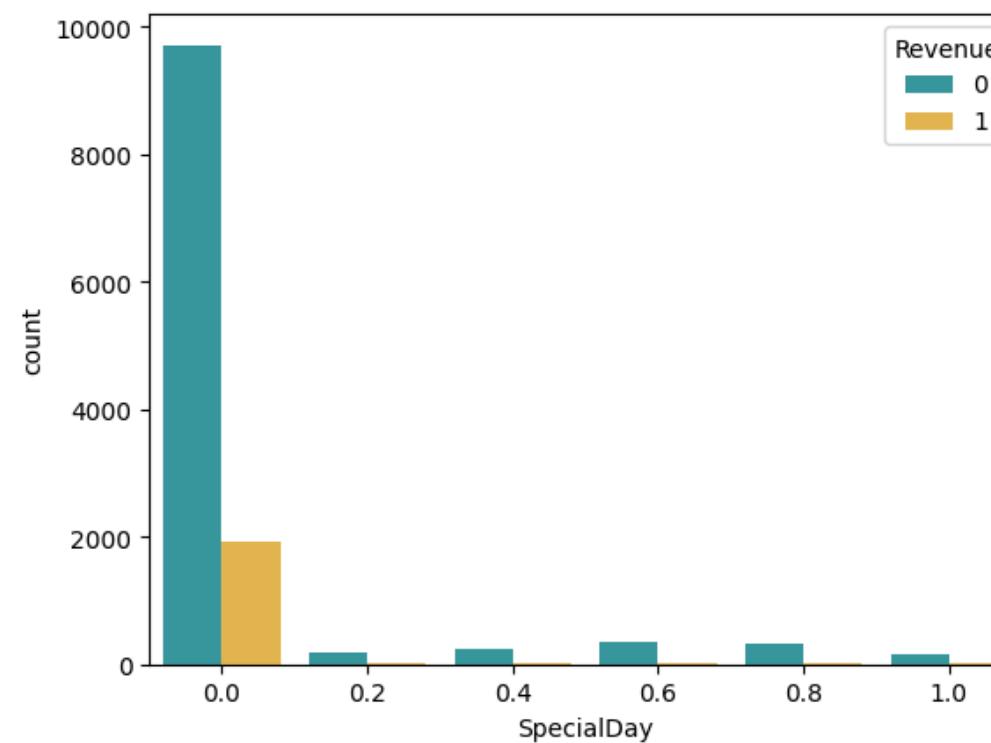
VISITOR TERKONVERSI BANYAK PADA WEEKEND

Revenue Konversi :

Weekend : 17,5%

Weekday : 14,9%

Promo khusus dan penawaran menarik diutamakan pada weekend



Special Day tidak berpengaruh terhadap peluang konversi

Perbaiki platform dan tingkatkan pengalaman pengguna

Data Preprocessing



Model 1:

1. Dilakukan drop terhadap data **duplicasi** data sebesar **5,5%** dari data keseluruhan.
2. Menghapus **seluruh missing values** pada semua fitur (administrative, administrative_duration, product_related_duration, dan operating_systems) dengan total persentase penghapusan adalah **0,14%** dari data duplikasi yang telah dihapus.

Model 2:

1. Menghapus duplikasi seperti **Model 1**
2. Menghapus *missing values* pada fitur **administrative** dan **bounce_rate**.
3. Mengasumsikan *missing values* pada **produk_related_duration** dan **administrative_duration** dengan jumlah dari masing-masing (**product_related** atau **administrative**) yang kemudian dikalikan 120 detik.
4. Mengisi *missing values* **operating_systems** dengan **modus**

Model 3:

Memberikan *treatment* seperti **Model 2** dan menghapus outlier **exit_rates** yang sebesar **0,08%** dari data setelah *treatment* **Model 2** dilakukan.

Model 5:

- Melakukan *treatment* seperti **Model 2**. Kemudian, melakukan **feature extraction** pada model ini dengan cara menambahkan beberapa fitur sebagai berikut.
1. total_visits
 2. total_duration
 3. visit_bounce_rates
 4. visit_exit_rates
 5. special_day_visits
 6. special_day_duration

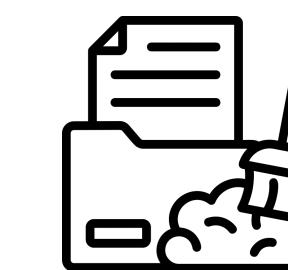
Model 6:

Memberikan perlakuan seperti **Model 3**, kemudian melakukan **oversampling** dengan SMOTE yang disertai beberapa perbandingan oversampling sebagai berikut.

1. Resampled 50:50
2. Resampled 50:40
3. Resampled 50:35
4. Resampled 50:30
5. Resampled 50:25
6. Resampled 50:20

Model 7:

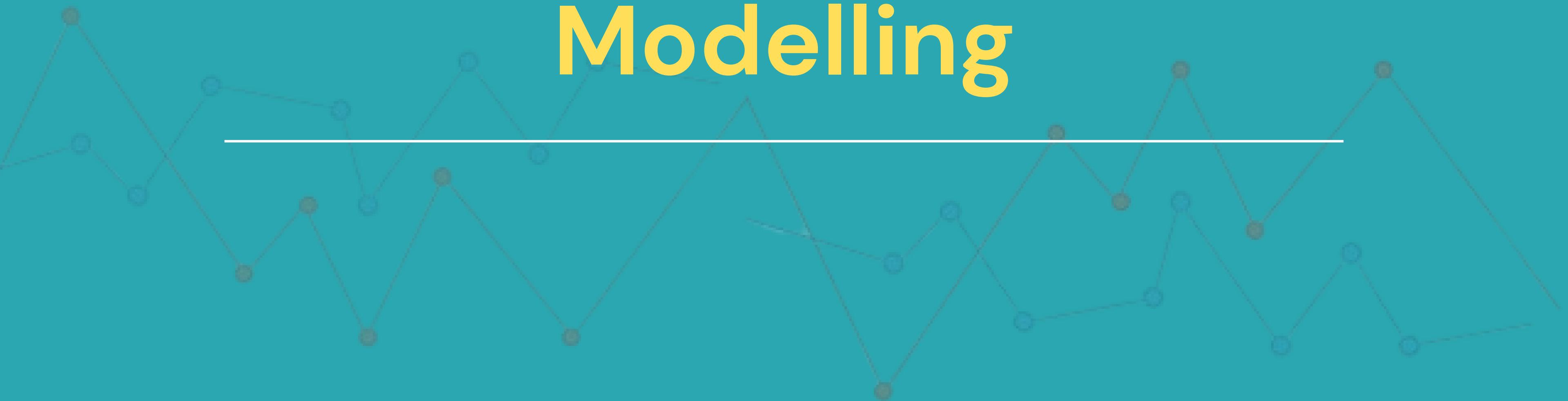
Selain melakukan perlakuan seperti model 6, dilakukan standarisasi data pada tiap x_train yang telah di-resample dan x_test



SKENARIO MODEL

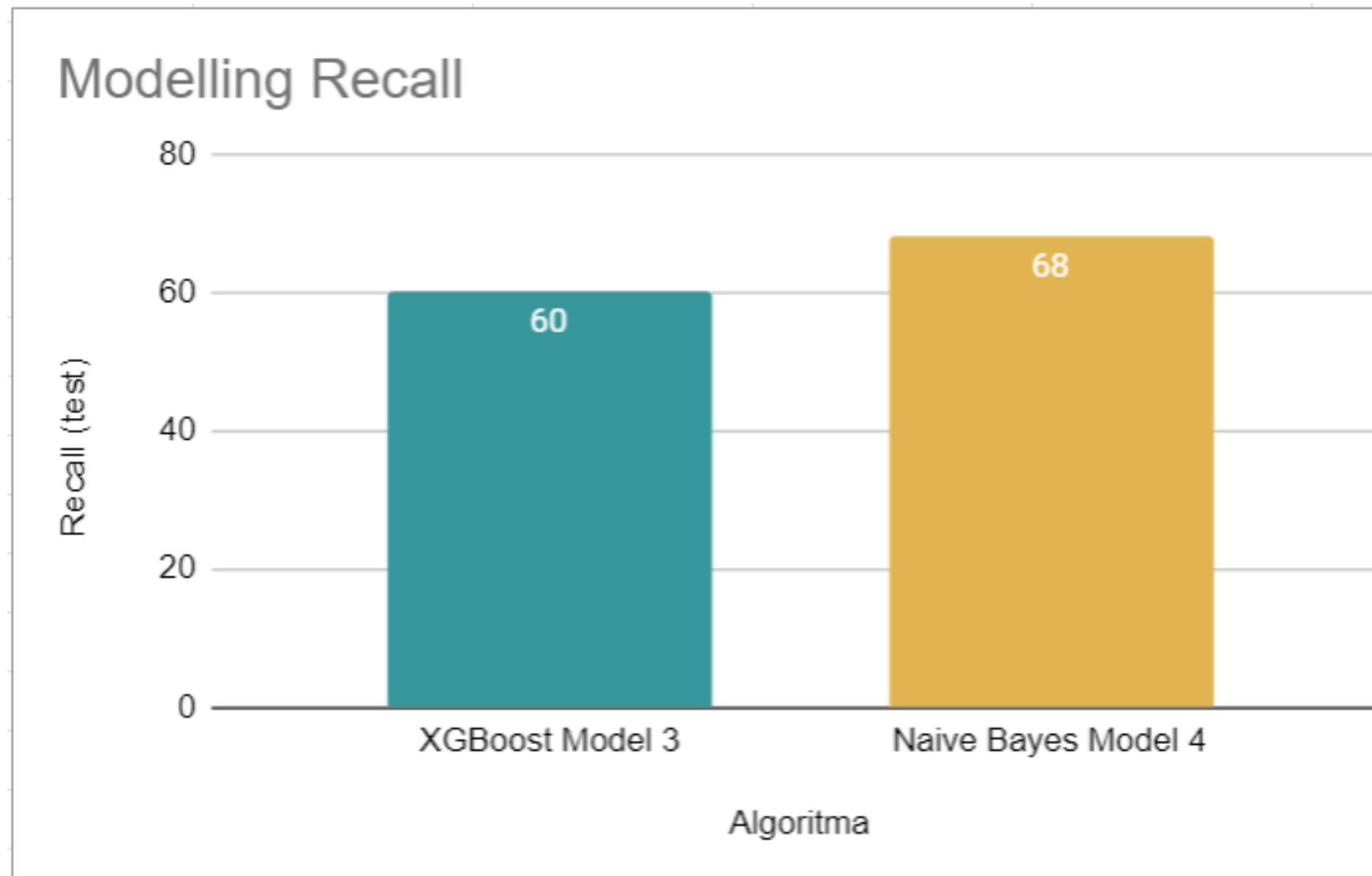
Memberikan perlakuan seperti **Model 2**. Kemudian, melakukan **standarisasi** data pada x_train dan x_test.

Modelling



Evaluasi metric yang difokuskan adalah **Recall** dengan tujuan untuk meminimalisir False Negative, sehingga **kita tidak kehilangan potensial pelanggan.**

Pemodelan dilakukan dengan berdasarkan dari data preprocessing dimana untuk melihat performa dari **7 skenario pemodelan agar tidak memilih hasil nilai yang overfit.**



$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Dari keseluruhan skenario pemodelan, **pemodelan 3 dengan XGBoost dan pemodelan 4 dengan Naive Bayes** memiliki potensial yang besar, dimana performanya cukup baik dibanding lainnya. -> Optimasi 2 model ini

RECALL SETELAH HYPERPARAMETER TUNNING

XG Boost

Naive Bayes

Parameter yang di tuning:

max_depth : [3, 5, 7]
learning_rate : [0.1, 0.01, 0.001]
subsample : [0.5, 0.7, 1]

Parameter yang di tuning:

priors : [None, [0.4, 0.6], [0.5, 0.5],[0.3, 0.7],[0.2,0.8],[0.1,0.9]]
var_smoothing : np.logspace(0, -9, num=100) }

59

97

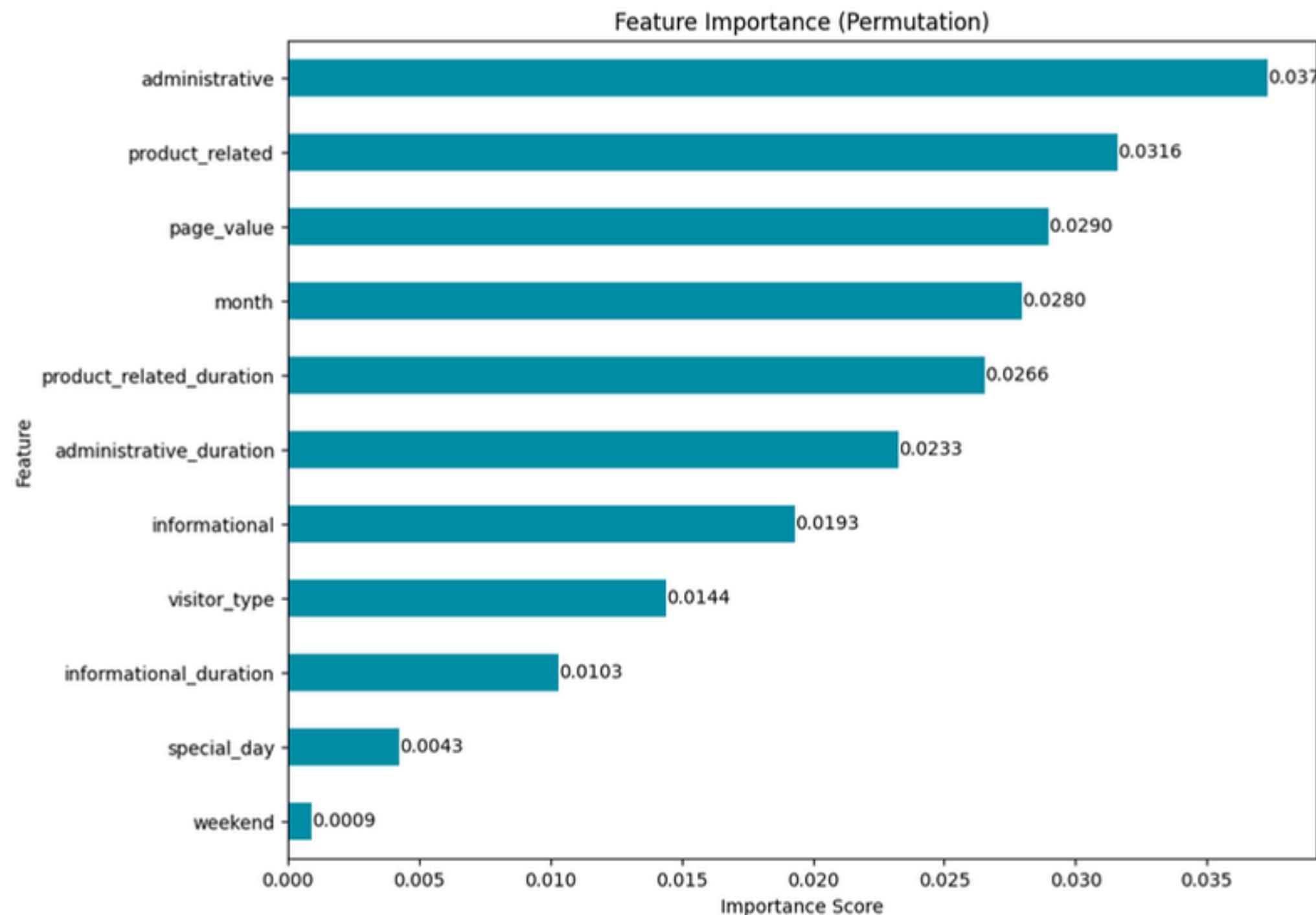
Recall (Test Set): 0.97, menunjukkan bahwa **model mampu mendekripsi 97% dari semua kasus konversi yang sebenarnya.**

Dengan recall setinggi ini, **model sangat dapat diandalkan untuk memastikan tidak ada konversi yang terlewatkan**, sesuai dengan prioritas utama model ini yaitu mengidentifikasi semua konversi.

Confusion Metrix Model Naive Bayes

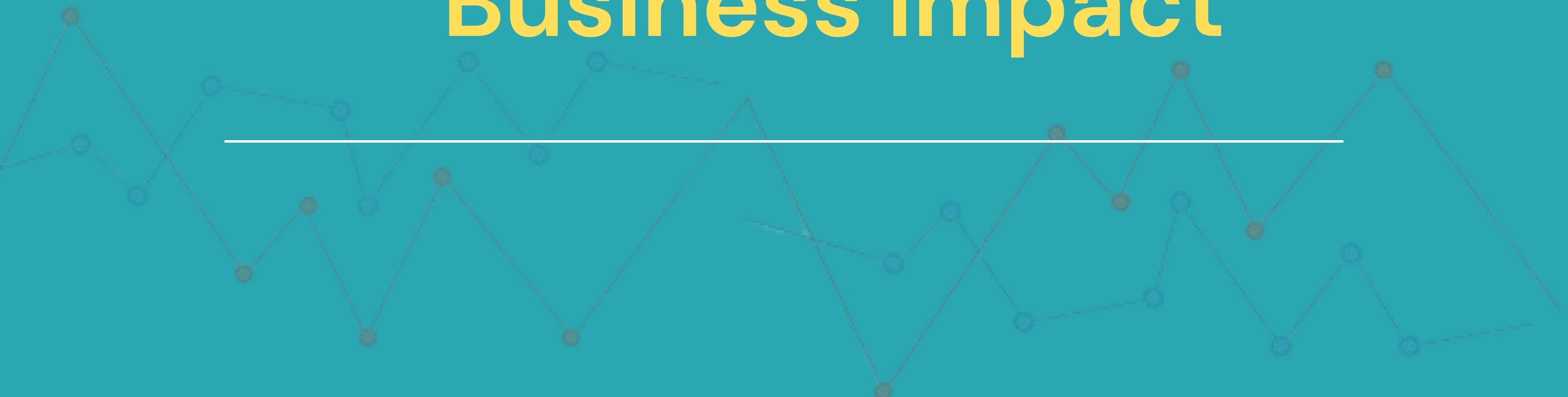
		Predicted	
		No	Yes
Actual	No	TN 2367	FP 7811
	Yes	FN 36	TP 1847

FEATURE IMPORTANCE



Dari Pemodelan Naive Bayes yang telah di tuning, terlihat **terdapat 4 fitur yang sangat penting dalam memprediksi apakah visitor akan terkonversi sebagai pelanggan atau tidak**, yaitu : **Administrative, product related, month, page value, dan product related duration.** Fitur-fitur ini memiliki dampak paling signifikan terhadap tingkat konversi pengguna, yang mendukung hasil analysis sebelumnya di EDA (pada heatmap correlation). **Maka sangat direkomendasikan perusahaan fokus mengoptimalkan fitur fitur tersebut.**

Business Impact



BUSINESS IMPACT

Perubahan Pada Tingkat Konversi(CVR)

- konversi raw data : **15,59%**
- **prediksi conversion rate model : 81,63%**
- **Peningkatan: $81,63\% - 15,59\% = 66,37\%$**

Jumlah Konversi dan Non-Konversi Baru

- Data aktual dari dataset, **12.061 pengunjung** (sudah tidak ada duplikat)
- **Konversi: 81,63%** dari 12.061 \approx 9.845 pengunjung
- **Non-Konversi: 18,37%** dari 12.061 \approx 2.216 pengunjung

Asumsi Pendapatan

- Misalkan **rata-rata pendapatan** per konversi adalah **\$50**
- **Sebelum** Penerapan Model : $1.880 \text{ pengunjung} \times \$50 = \$94.000$
- **Sesudah** Penerapan Model : $9.845 \text{ pengunjung} \times \$50 = \$492.250$
- **Peningkatan Pengunjung** : $9.845 - 1.880 = 7965 \text{ pengunjung}$
- **Peningkatan Revenue** : $\$492.250 - \$94.000 = \$398.250$

Recommendation

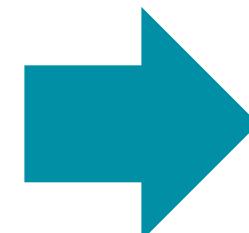


01

Mengidentifikasi
faktor yang
mempengaruhi
konversi pelanggan

FEATURE IMPORTANCE

- 1.administrative
- 2.product_related
- 3.page_value
- 4.month
- 5.product_related_duration
- 6.administrative_duration
- 7.informational
- 8.visitor_type
- 9.informational_duration
- 10.special_day
- 11.weekend



Diperlukan fokus
mengoptimalkan **5 fitur**
teratas untuk
meningkatkan
conversion rate.

02

Meningkatkan penjualan perusahaan

Insight

- Fitur **Page Value**, **Product Related**, dan **Administrative** secara signifikan meningkatkan peluang **pengunjung untuk melakukan pembelian**.
- Returning visitor mendominasi traffic website dengan konversi pengunjung yang rendah menandakan sebagian **visitor senang menjelajahi produk**, tetapi **produk yang ditemui belum sesuai dengan preferensi mereka**.
- Terdapatnya peningkatan konversi pada bulan **Mei** dan **November**, serta pada waktu **weekend**.

Rekomendasi

- Fokus pada peningkatan **Page Value** dan **Product Related** dengan menyediakan konten yang relevan dan menarik untuk mendorong konversi.
 - Optimasi deskripsi produk
 - Manfaatkan data perusahaan (aktivitas visitor) untuk mengimplementasikan rekomendasi produk sesuai dengan preferensi setiap visitor.
- Manfaatkan momen **akhir pekan**, bulan mendekati bulan **Mei** dan **November** (bulan menuju hari raya atau akhir tahun) untuk mengoptimalkan pemasaran.

03

Mengantisipasi penurunan penjualan lebih lanjut

Menerapkan model prediksi dengan baik
dan melakukan evaluasi, serta
pemantauan secara berkala terhadap
performa kampanye dan model

THANK YOU



APPENDIX

Model	Algoritma	Accuracy (test)	Accuracy (train)	Recall (test)	roc_auc (test)	roc_auc (train)	crossval (train)	crossval (test)	CV	
1	XGBoost	90	99	60	92	100	94	56	89	Overfit
	ETC	90	100	50	92	100	100	47	89	
	Naive Bayes	84	85	53	83	84	53	50	84	
2	XGBoost	90	99	61	93	100	92	57	89	Overfit
	ETC	90	100	53	92	100	100	48	89	
	Naive Bayes	85	84	53	85	83	52	49	84	
3	XGBoost	86	99	60	91	100	95	60	87	
	ETC	85	100	47	90	100	100	50	87	
	Naive Bayes	81	82	40	78	81	44	42	81	
4	XGBoost	90	99	61	93	100	95	60	89	
	ETC	87	100	56	74	100	100	56	86	
	Naive Bayes	82	81	68	84	82	46	43	80	
5	XGBoost	89	99	60	93	100	94	56	89	Overfit
	ETC	90	100	46	92	100	100	42	88	
	Naive Bayes	84	83	59	84	82	58	55	82	
6	XGBoost									
	50:50:00	89	99	65	92	100	92	56	91	
	50:40:00	89	99	64	92	100	92	56	91	
	50:35:00	89	99	65	92	100	92	56	91	
	50:30:00	89	99	62	92	100	92	56	91	
	50:20:00	89	99	66	92	100	92	56	90	Overfit
	50:25:00	89	99	62	92	100	92	56	90	
	ETC									
	50:50:00	62	100	89	83	100	100	100	47	93
	50:40:00	73	100	85	86	100	100	100	46	92
	50:35:00	76	100	81	86	100	100	100	47	92
	50:30:00	85	100	71	89	100	100	100	47	91
	50:20:00	90	84	59	85	82	92	56	91	
	50:25:00	89	100	63	92	100	100	100	47	91
	Naive Bayes									
7	Naive Bayes									
	50:50:00	48	73	53	50	85	66	64	73	
	50:40:00	51	72	50	50	85	66	64	72	
	50:35:00	53	72	48	50	85	66	64	72	
	50:30:00	54	72	44	50	84	66	64	71	
	50:20:00	60	74	35	50	84	66	64	73	
	50:25:00	56	72	40	50	84	66	64	72	

Naive Bayes	50:50:00	80	78	66	81	86	52	49	77	
	50:40:00	81	78	65	81	85	52	49	77	

Hasil pada percobaan modeling dengan penggunaan 3 model terbaik, dapat dilihat terhadap nilai recall yang lebih tinggi tetapi cenderung overfit