Online Shoppers Purchase Intention Dataset

LAMBDASHOP

LAMBDASHOP

Tim
Data Scientist

Rekomendasi Bisnis untuk

Revenue

Cost

dengan cara
Membuat kriteria pengunjung online shopping

OUR TEAM

Iqbal

Shirley

Rahma

Syofwan

Ghofur

Dzaky

Tohar

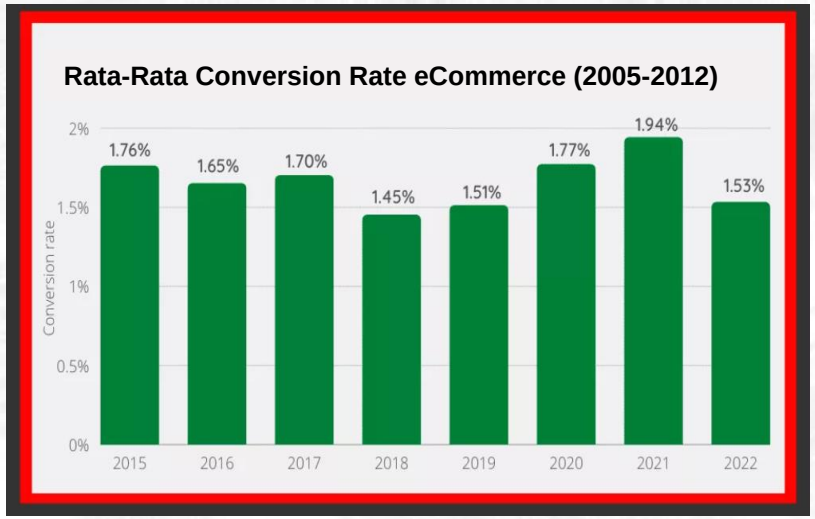
Widi

OUTLINE

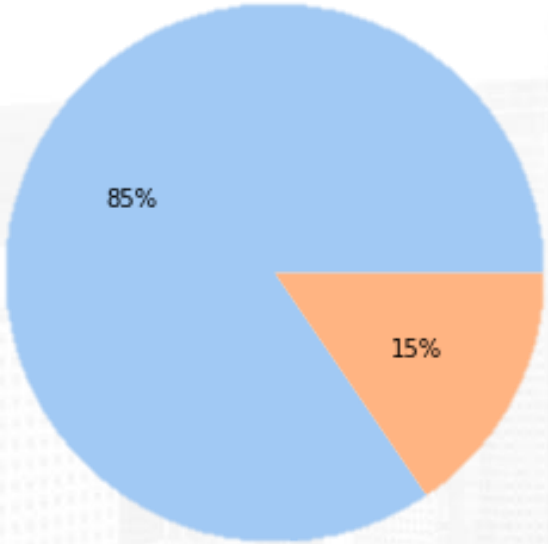
- Latar Belakang Masalah
- Goal, Objective, Business Metrics
- EDA
- Data Pre-Processing, Modelling
- Rekomendasi Bisnis

Latar Belakang Masalah

Menurut hasil survei IpsosGlobal Trend 2021. **Di Indonesia, 73% konsumen memilih belanja online ketimbang belanja di toko dengan pertimbangan lebih mudah**.



Terdapat **penurunan rata-rata efektivitas halaman bisnis** untuk menarik pengunjung melakukan sebuah tindakan.



Pada saat ini, di **LAMBDASHOP** hanya **15%** dari visitor platform e-commerce kami berakhir pada transaksi yang menghasilkan revenue.

Hal ini mendorong kami untuk menganalisa **pola-pola ketertarikan dan perilaku customer dalam berbelanja** sehingga bisa membuat model yang tepat untuk memprediksi kecenderungan konsumen untuk melakukan transaksi dan **menghasilkan revenue dan menurunkan marketing cost atau promotion cost.**

GOAL

- Memprediksi user yang dapat menghasilkan revenue / potential user
- Menentukan fitur apa yang paling mempengaruhi suatu user untuk melakukan transaksi
- Memberikan hasil Analisa kepada tim bisnis, agar mereka dapat membuat keputusan terbaik bagaimana cara meningkatkan jumlah user yang melakukan transaksi dan menurunkan marketing cost atau promotion cost

OBJECTIVE

Membuat model **Machine Learning** yang dapat memprediksi user yang memiliki potensi untuk melakukan transaksi, dengan cara membuat segmentasi user mendetail seperti ***'Beli', dan 'Tidak Beli'***. Segmentasi ini diharapkan menjadi bahan pertimbangan tim business agar lebih efektif dan efisien dalam mengelola marketing cost atau promtion cost.

BUSINESS METRICS

- Meningkatkan Conversion Rate Pengunjung LambdaShop
- Mengoptimalkan marketing cost atau promtion cost agar efektif dan efisiensi

EDA

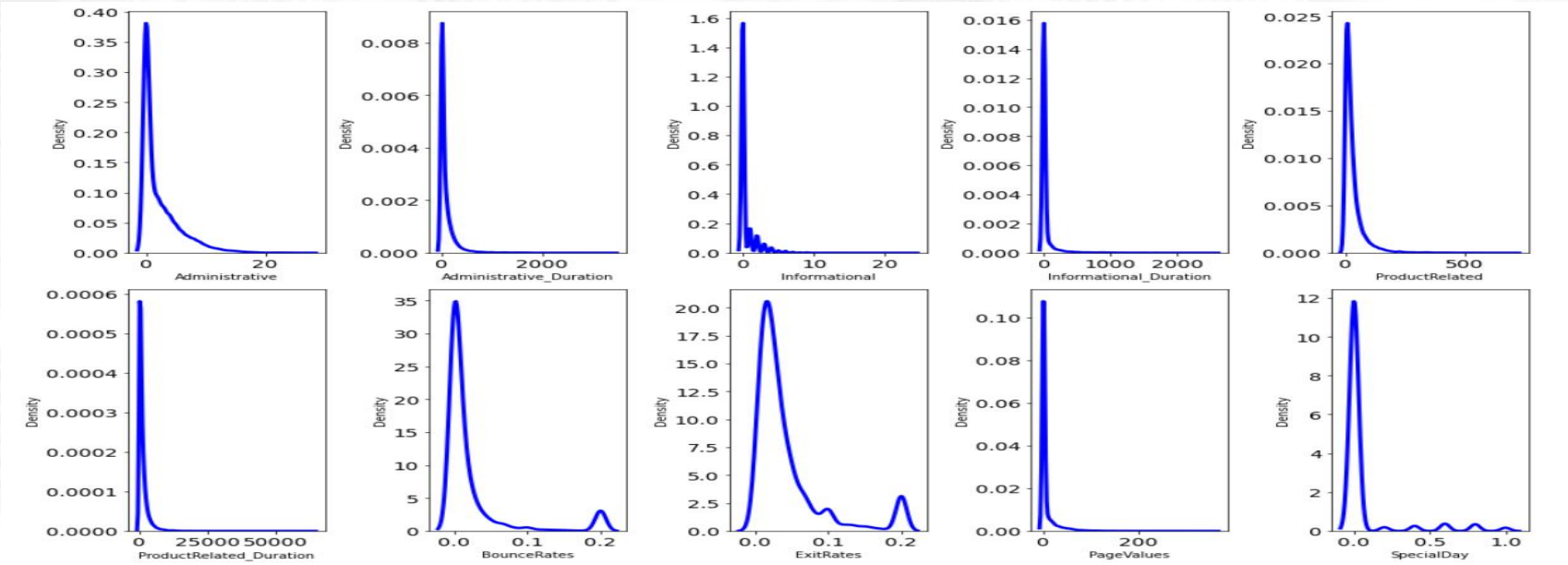
Exploratory Data Analysis

Rakamin Academy

Exploratory Data Analysis

Analisa Deskriptif

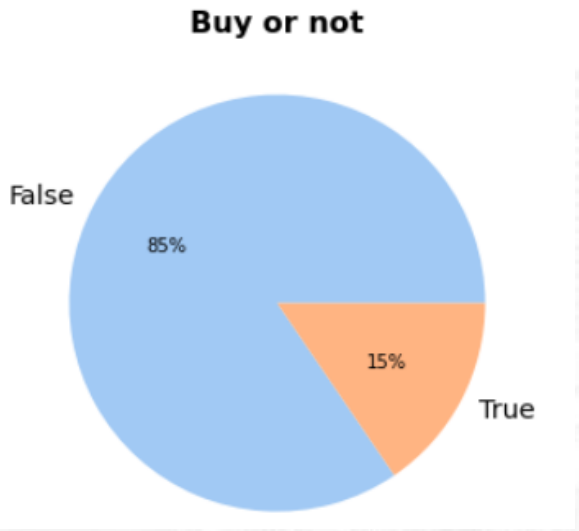
- Data Set memiliki **12.330** row data
- Tidak terdapat kolom yang memiliki nilai kosong
- Terdapat **125 nilai duplikat**
- Data relatif memiliki atribut bernilai min/max terlalu jauh dari mean/median sehingga termasuk **positive skewed distribution.**

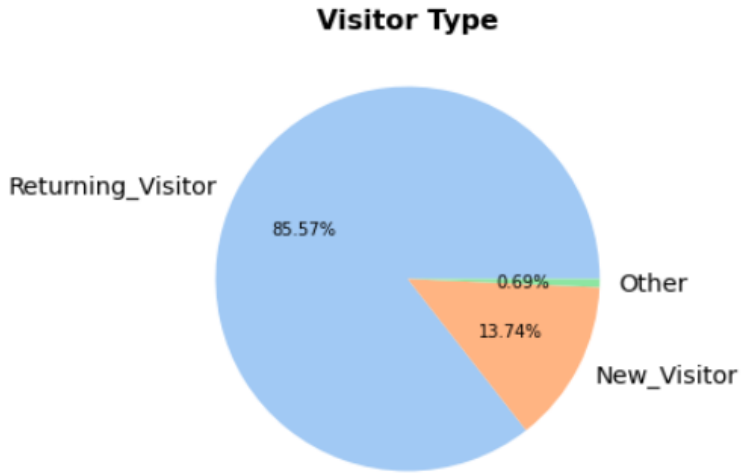


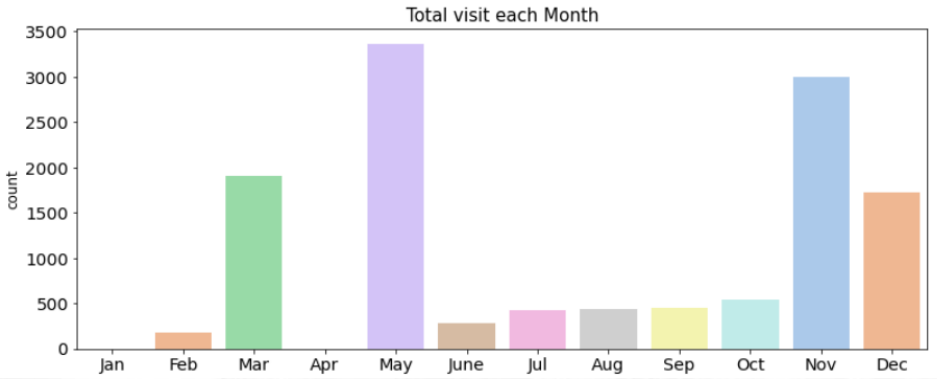
Numerical data	Categorical data
Administrative	OperatingSystems
Administrative_Duration	Browser
Informational	Region
Informational_Duration	TrafficType
ProductRelated	Month
ProductRelated_Duration	VisitorType
BounceRates	Weekend
ExitRates	Revenue
PageValues	
SpecialDay	

Exploratory Data Analysis

Analisa Diskriptif





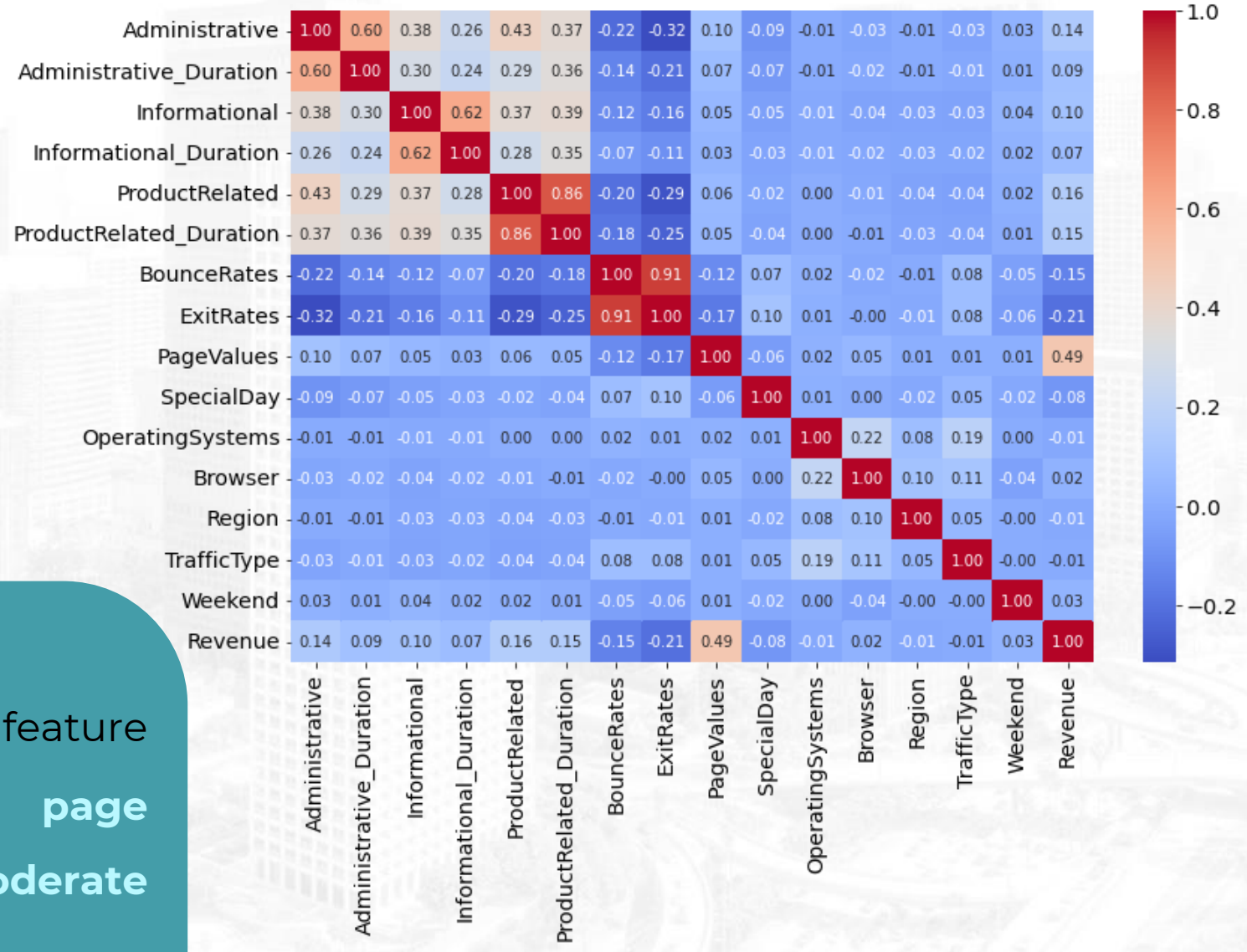


Univariate Analytics

- Terdapat **banyak outlier** untuk fitur-fitur numerik.
- Terdapat class imbalance pada target variabel "Revenue".
- Nilai dominan dari masing-masing fitur sebagai berikut:
  - → Revenue - Dari 12.330 sesi dalam dataset, diketahui **84,5% (10.422)** adalah data visit yang tidak menghasilkan revenue dan hanya **15,5% (1.908)** visitor yang menghasilkanrevenue.
  - → Weekend- Visitor lebih banyak di Weekdays daripada di weekend
  - → Month - Bulan dengan jumlah visitor signifikan ada di **Mei, November, Maret** dan **Desember**. Terdapat 2 bulan tanpa visitor yaitu bulan Januari dan April.
  - → Visitor Type - Visitor yang dominan merupakan **Returning Visitor**.

Pre Processing

Visualisasi Data



Berdasarkan nilai korelasi antara feature dan target **Revenue**, feature **page values mempunyai korelasi moderate positif dengan target.**

Rakamin Academy

PRE-PROCESSING

Pre Processing

Data Cleansing

A. Handle duplicated data

Menghapus data duplikat sebanyak 125 data **dari 12330 menjadi 12205**.

B. Handle outliers

- Menggunakan **Zscore** menghasilkan perubahan jumlah data **dari 12205 menjadi 10020. (>10% data)**
- Menggunakan **IQR** dan **Flooring and Capping dari 12005 menjadi 5150** terlalu banyak data yang dihapus.

Oleh Karena itu **tidak perlu dilakukan handling outlier menggunakan metode ini.**

C. Feature transformation

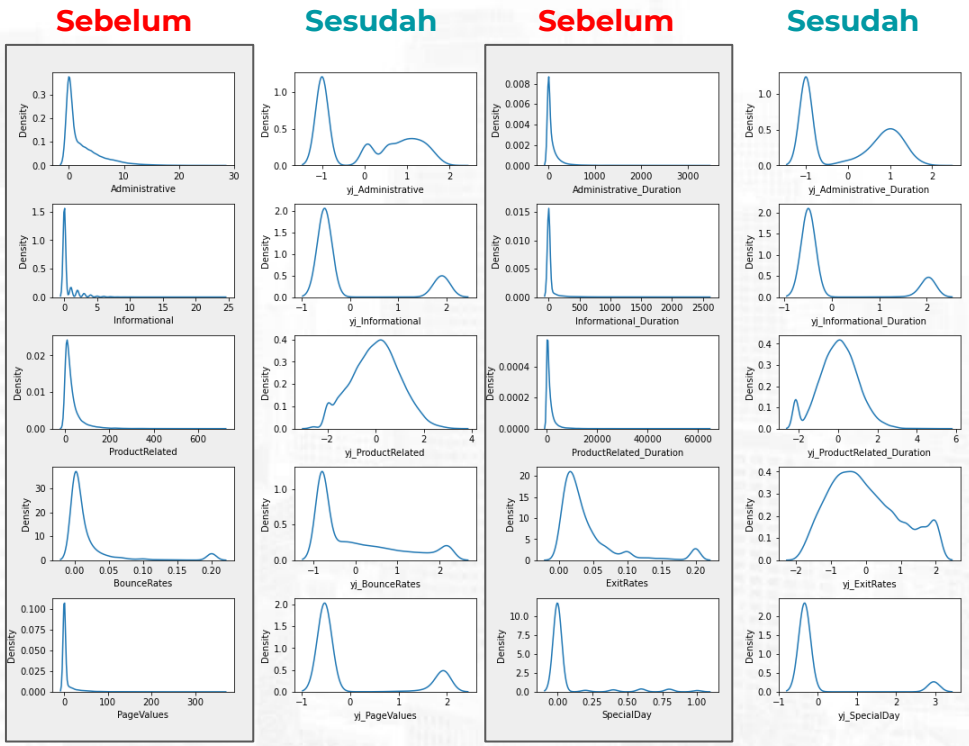
Menggunakan **Yeo-Jonshon -** Data awal yang positive skewed setelah di transformasi menjadi **mendekati normal skewed**

D. Feature Encoding

- **One Hot Encoding** untuk fitur VisitorType
- **Label Encoding** untuk Fitur Weekend, Month (ordinal)
- **Encoding Threshold** untuk fiture OS, browser type, traffic type, Region

E. Feature Class Imbalance

Menggunakan **SMOTE** pembagian False/True pada rasio 5:3



Pre Processing

Feature Engineering

Feature Selection

Dari Analisa diatas maka disimpulkan bahwa :

- Dari hasil **ANOVA f-test** Feature Selection terlihat bahwa feature **PageValue** adalah fitur yang paling relevan.
- Feature numerik yang akan di **drop** adalah : **BounceRates.**.

  Feature **ExitRates** berkorelasi dengan **BounceRates** dengan nilai diatas 0.7

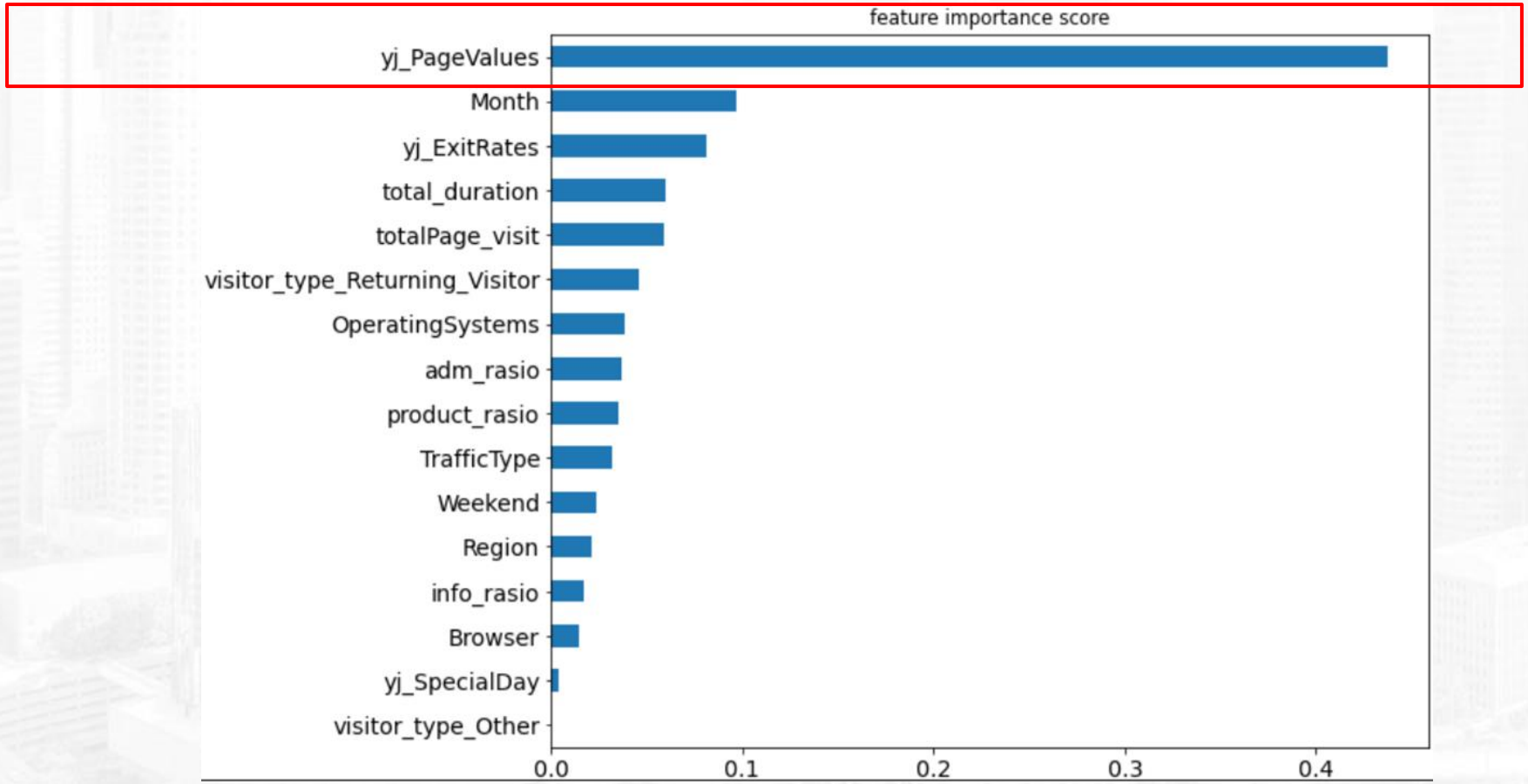
  sehingga masuk kategori **redundant feature**
- Melakukan **kombinasi** antar fitur yang mempunyai korelasi kuat lainnya untuk menghindari pengulangan.

Feature Extraction

- **Total Duration** merupakan Administrative_Duration + Informational_Duration + ProductRelated_Duration
- **Total Visit** merupakan Total jumlah page yang dikunjungi per masing-masing sesi
- **Rasio (Duration/Page)**
  - adm_rasio = Administrative_Duration/ Administrative
  - info_rasio = Informational_Duration / Informational
  - product_rasio = ProductRelated_Duration / ProductRelated

Modelling Experiments

Feature Important



Dari grafik tersebut diketahui bahwa **Page Value** adalah memberikan yang memberikan dampak paling besar

Rakamin Academy

MODELLING EXPERIMENT

Modelling Experiments

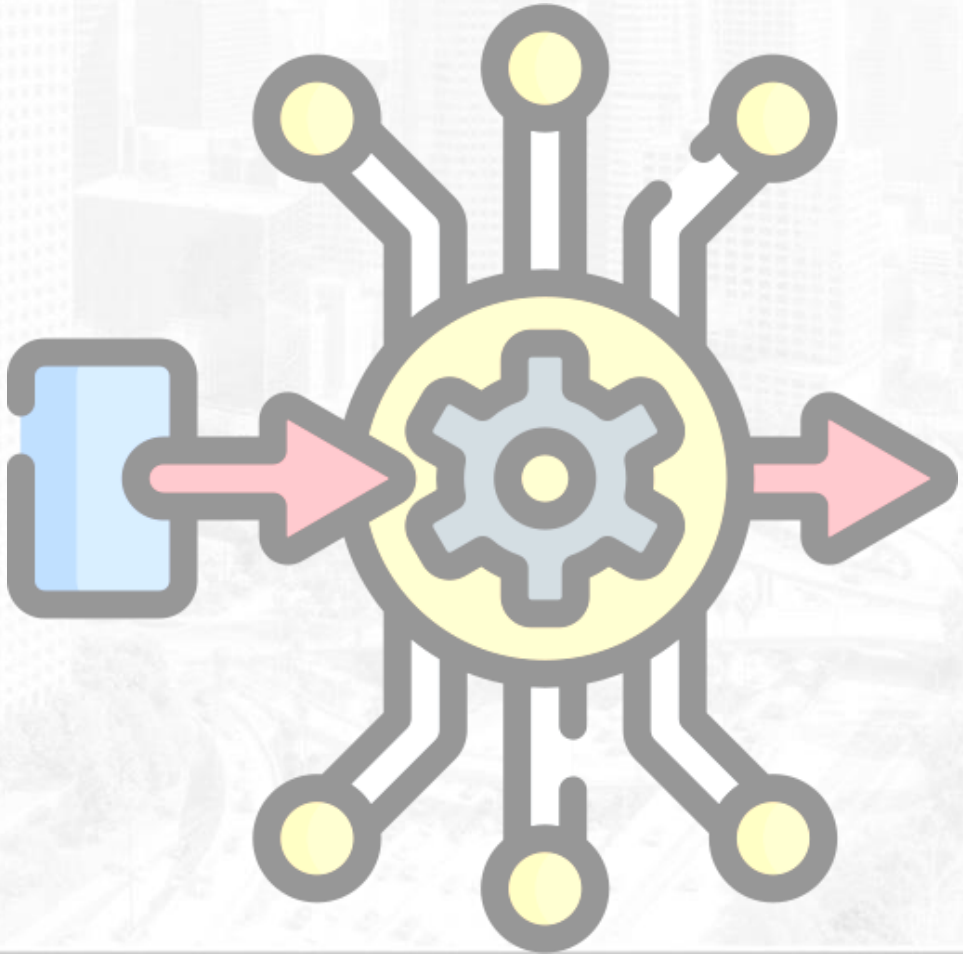
split data train dan data test dengan **rasio perbandingan 80:20**

Model

- **Logistic Regression**
- **K-Nearest Neighbor**
- **Decision Tree**
- **Random Forest**
- **Adaboost**
- **XGBoost**

Hyperparameter Tuning

- Seluruh modelling



Modelling Experiments

Split dan Modelling

Hasil

Fit	Logistic Regression	K-Nearest Neighbor	Decision Tree	Random Forest	Adaboost	XGBoost
Accuracy(Test Set)	0,88	0,82	0,86	0,89	0,88	0,88
Precision (Test Set)	0,6	0,47	0,57	0,68	0,63	0,65
Recall (Test Set)	0,8	0,7	0,61	0,71	0,69	0,7
F1-Score (Test Set)	0,68	0,56	0,59	0,69	0,66	0,67
ROC AUC (Test-Proba)	0,91	0,85	0,76	0,93	0,91	0,93
ROC AUC (Train-Proba)	0,91	0,99	1	1	0,96	0,98
ROC AUC (Crossval-train)	0,9	0,95	1	1	0,92	0,94
ROC AUC (Crossval-test)	0,88	0,82	0,712	0,9	0,89	0,91

Hyperparameter Tuning	Logistic Regression	K-Nearest Neighbor	Decision Tree	Random Forest	Adaboost	XGBoost
Accuracy(Test Set)	0,88	0,87	0,85	0,89	0,88	0,89
Precision (Test Set)	0,6	0,48	0,57	0,66	0,62	0,68
Recall (Test Set)	0,8	0,56	0,6	0,75	0,69	0,68
F1-Score (Test Set)	0,68	0,52	0,58	0,7	0,66	0,68
ROC AUC (Test-Proba)	0,91	0,72	0,76	0,93	0,92	0,93
ROC AUC (Train-Proba)	0,91	1	1	0,99	0,96	1
ROC AUC (Crossval-train)	0,9	0,93	0,92	0,96		0,99
ROC AUC (Crossval-test)	0,88	0,87	0,86	0,9		0,9

Dari hasil modeling yang telah dilakukan, **Random Forest Modelling yang telah di tuning** yang menjadi pilihan kami dengan **Akurasi 0.89 dan Recall 0.75**

Impact Model

Simulasi Awal

Total Customer = 2441 orang

Conversion Rate = 15%

cost = Rp. 10.000 / orang
revenue = Rp. 100.000 / orang

Total Cost
= Rp. 10.000 x (2441)
= Rp. 24.410.000

Total Revenue
= Rp.100.000 x (2441*15%)
= Rp.100.000 x 366.15
= Rp.36.615.000

Nett
= Revenue - Cost
= Rp.36.615.000 - Rp. 24.410.000
= Rp.12.205.000

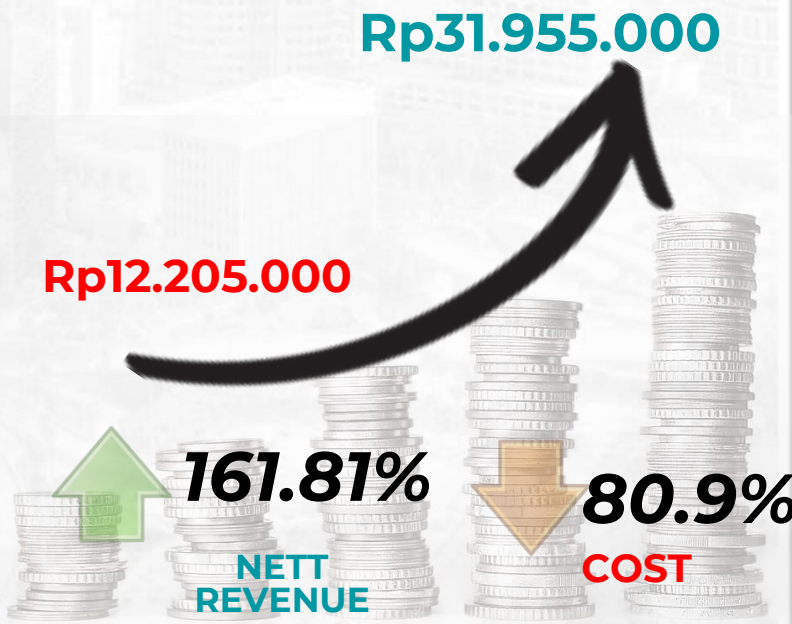
		Prediksi: Beli	Prediksi: Tidak Beli
Actual	Beli	307	105
	Tidak Beli	159	1870

Simulasi Impact Model

Total Cost (jumlah yang terdeteksi membeli)
= Rp.10.000 x (TP+FP)
= Rp.10.000 x (307+159) **= Rp.4.660.000**

Total Revenue
= Rp.100.000 x (2441*15%)
= Rp.100.000 x 366.15
= Rp.36.615.000

Nett
= Revenue - Cost
= Rp.36.615.000 - Rp. 4.660.000
= Rp.31.995.000



REKOMENDASI BISNIS

Rekomendasi Bisnis

1. Page value paling berkorelasi dengan revenue, maka page **dengan nilai page value terendah perlu dioptimalisasi**

2. Tim Bisnis dan marketing perlu membuat **promosi yang sesuai berdasarkan bulan-bulan** tertentu karena terdapat bulan yang memiliki tinggi visitor dan ada pula yang bahkan tidak ada pengunjung.

3. Tim bisnis disarankan lebih **fokus mengelola pengunjung lama daripada pengunjung baru**, karena peluang revenue lebih besar dan menekan cost promotion

4. Perlu ditambahkan feature-feature baru guna menaikkan nilai Page Value seperti : **Customer ID , Gender , Seller Reputation , Satisfaction Score dan Product in Basket**

Rakamin Academy

TERIMA KASIH