Soal Mentoring Probability

Job Preparation Program, Data Science, Pacmann AI

Catatan:

- Make a copy docs ini sebelum menjawab.
- Ada dua fokus dalam menjawab soal ini
  - Menemukan jawaban yang benar
  - Memahami cara mendapatkan jawaban yang benar
- Selama mengerjakan soal, Anda **wajib melampirkan** cara & proses menjawab. **BUKAN CUMA** pasang rumus, tapi **berikan elaborasi** kenapa Anda melakukan hal tersebut.
- Silahkan lampirkan proses Anda dengan menuliskan langsung pada docs atau melampirkan foto proses (yang dapat dibaca).
- Kumpulkan docs Anda sesuai dengan link submission yang tersedia.
- Selamat mengerjakan 🫡

---

1. [10 points] Airworthiness adalah salah satu hal penting yang dijadikan acuan sebuah pesawat dapat terbang atau tidak. Saat proses maintenance pesawat, akan dicek komponen yang malfungsi dan nantinya dipanggil teknisi untuk membantu memperbaikinya.

   Asumsikan **X** adalah bilangan random yang mendefinisikan jumlah komponen yang malfungsi dan **Y** adalah bilangan random yang mendefinisikan berapa kali teknisi dipanggil. Berikut adalah kombinasi kejadian **X** dan **Y**

	X = x			
Y = y	**0**	**1**	**2**	**3**
0	15	30	5	0
1	5	15	5	5
2	0	5	10	5

   Pertanyaannya

   a. [2 points] Pesawat dikatakan tidak airworthy ketika jumlah komponen yang malfungsi lebih dari 1. Berapa peluang pesawat tidak airworthy berdasarkan tabel di atas?

1. Ubah bentuk data menjadi joint table berisi joint PMF jumlah komponen (X) dan jumlah berapa kali teknisi dipanggil (Y). Setiap data jumlah komponen dan jumlah kali teknisi dipanggil dibagi dengan penjumlahan keseluruhan data di sample space.

	X = x			
Y = y	0	1	2	3
0	0,15	0,3	0,05	0
1	0,05	0,15	0,05	0,05
2	0	0,05	0,1	0,05

2. Hitung Marginal PMF atau peluang komponen malfungsi > 1 = P(X>1) untuk semua nilai Y

P(X>1) = P(X=2))U P(X=3)
P(X>1) = P(X=2, Y=0) + P(X=2, Y=1) + P(X=2, Y=2) + P(X=3,Y=0) + P(X=3, Y=1) + PX=3, Y=2)
P(X>1) = (0.05 + 0.05 + 0.1) + (0 + 0.05 + 0.05)
= 0.2 + 0.1
= 0.3
Jadi, peluang pesawat tidak airworthy adalah 0.3 atu 30%

b. [3 points] Cari peluang teknisi dipanggil

Hitung Marginal PMF atau peluang teknisi dipanggil = P(Y>0) untuk semua nilai X
P(Y>0) = P(Y=1) U P(Y=2)
= P(Y=1) + P(Y=2)
= (0.15 + 0.05 + 0.05) + (0.05 + 0.1 + 0.05)
= 0.25 + 0.25
= 0.5
Jadi, peluang teknisi dipanggil adalah 0.5 atau 50%

c. [2 points] Apabila diketahui satu teknisi spesifik memperbaiki satu komponen malfungsi dan pemilik pesawat akan rugi apabila ada lebih banyak teknisi datang dibanding jumlah komponen yang malfungsi, berapa peluang pemilik pesawat mengalami kerugian tersebut?
Tulis jawaban disini
Hitung peluang pesawat rugi = P(Y>X)
Buat tabel frekuensi Y>X
P(Y>X) = P(Y=1, X=0) + P(Y=2, X =0) + P(Y=2, X=1)
= 0.05 + 0 + 0.05
= 0.1
Jadi, peluang pesawat rugi adalah 0.1 atau 10%

d. [3 points] Periksa apakah kejadian random **X** dan **Y** independen.

X dan Y independen jika P(X=x, Y=y) = P(X=x) x P(Y=y)
Karena kalau dua kejadian independen, maka peluang gabungannya **tidak ada interaksi** antar kejadian.
Contoh pada tabel peluang berikut:

	X = x			
Y = y	0	1	2	3
0	0,15	0,3	0,05	0
1	0,05	0,15	0,05	0,05
2	0	0,05	0,1	0,05

P(X=0, Y=0) - 0.15
P(X=0) = 0.15 + 0.05 + 0 = 0.2
P(Y=0) = 0.15 + 0.3 + 0.05 + 0 = 0.5
P(X=0) x P(Y=0) = 0.2 x 0.5 = 0.1

Jadi, P(X=0, Y=0) tidak sama dengan P(X=0) x P(Y=0)
X dan Y tidak independen.

2. [27 points] Anda bekerja sebagai marketing analyst di industri FMCG. Sebagai marketing analyst, Anda dituntut untuk bisa mengoptimasi proses marketing agar targetan revenue tercapai. Anda biasa menggunakan ads untuk membantu penjualan. Diberikan data historis ads yang pernah Anda lakukan dalam 4 tahun terakhir sebagai berikut (lihat data: ads_data.csv)

	ads_id	date	impressions	leads
0	1923	2020-03-01 16:58:07	263620	3421
1	1924	2020-03-01 18:47:41	260940	6610
2	1925	2020-03-01 19:37:20	214612	6539
3	1926	2020-03-03 06:56:15	148663	8208
4	1927	2020-03-04 03:30:36	969414	4064

Rincian data:
- ads_id : ID ads
- date : Tanggal ads dimulai
- impressions : Impresi yang didapat dari ads setelah sehari diluncurkan
- leads : Leads yang didapat dari ads tersebut setelah sehari diluncurkan

Buatlah rencana marketing sederhana apabila target revenue Anda sebesar Rp 300.000.000.

Untuk menyelesaikannya, jawablah beberapa pertanyaan di bawah ini:

a. [3 points] Untuk menyederhanakan masalah, kategorikan data impresi & leads ke dalam kategori low, medium, dan high sesuai kriteria berikut:
   - Impresi
     - Kategori low: Impresi < 200.000,
     - Kategori medium: 200.000 <= impresi <= 600.000,
     - Kategori high: impresi > 600.000
   - Leads
     - Kategori low: Leads < 2.000,
     - Kategori medium: 2.000 <= leads <= 6.000,
     - Kategori high: leads > 6.000

   Kemudian buatlah pivot table untuk mencari jumlah ads untuk masing-masing pasangan kategori impresi - leads (contoh)

		Leads category		
		Low	**Medium**	**High**
Impressions category	**Low**			
	Medium			
	High			

1. Contoh hasil data yang telah dikategorikan impresi dan leads nya

ads_id	date	impressio	leads	imp_category	leads_category
1923	2020-03-0	263620	3421	Medium	Medium
1924	2020-03-0	260940	6610	Medium	High
1925	2020-03-0	214612	6539	Medium	High
1926	2020-03-0	148663	8208	Low	High
1927	2020-03-0	969414	4064	High	Medium
1928	2020-03-0	635885	6318	High	High
1929	2020-03-0	438614	9962	Medium	High
1930	2020-03-0	370442	6059	Medium	High
1931	2020-03-0	79701	9856	Low	High
1932	2020-03-0	509834	1071	Medium	Low
1933	2020-03-1	396877	2902	Medium	Medium
1934	2020-03-1	511448	1248	Medium	Low
1935	2020-03-1	474250	4978	Medium	Medium
1936	2020-03-1	416028	4143	Medium	Medium
1937	2020-03-1	6521	4078	Low	Medium
1938	2020-03-1	614579	7351	High	High
1939	2020-03-1	374802	4832	Medium	Medium
1940	2020-03-1	568124	1248	Medium	Low
1941	2020-03-1	514240	4683	Medium	Medium
1942	2020-03-2	572363	1830	Medium	Low
1943	2020-03-2	159719	2693	Low	Medium
1944	2020-03-2	462834	3179	Medium	Medium
1945	2020-03-2	382742	5898	Medium	Medium
1946	2020-03-2	494103	5782	Medium	Medium
1947	2020-03-2	226618	5709	Medium	Medium
1948	2020-03-2	847957	1522	High	Low
1949	2020-03-2	475307	2721	Medium	Medium

2. Dari data tersebut dibuat pivot table yang memuat count ads berdasarkan kategori impresi dan leads. Hasil pivot table adalah sebagai berikut:

Count of ads_id	Leads			
Impression	High	Low	Medium	Grand Total
High	188	83	179	450
Low	61	24	65	150
Medium	351	193	356	900
Grand Total	600	300	600	1500

b. [3 points] Kita bisa menganggap kategori impresi dan leads adalah random variabel diskrit. Dari pivot table sebelumnya, buatlah tabel yang berisi joint pmf untuk pasangan kategori impresi dan leads

Dari pivot table tersebut buat join table yang memuat joint PMF antara leads dan impressions dengan cara membagi jumlah tiap kategori dengan total ads.

Joint PMF Ads	Leads			
Impression	High	Low	Medium	Grand Total
High	0,125	0,1	0,119	0,299
Low	0,041	0	0,043	0,1
Medium	0,234	0,1	0,237	0,6
Grand Total	0,4	0,2	0,399	1

Contoh perhitungan:
P(Imp = High, Leads = High) = 188/1500 = 0.1253

Count of ads_id	Leads			
Impression	High	Low	Medium	Grand Total
High	188	83	179	450
Low	61	24	65	150
Medium	351	193	356	900
Grand Total	600	300	600	1500

Joint PMF Ads	Leads			
Impression	High	Low	Medium	Grand Total
High	=ROUND((B13/E16);3)			
Low				0,1
Medium	0,234	0,1	0,237	0,6
Grand Total	0,4	0,2	0,399	1

ROUND(number; num_digits)

c. [3 points] Berapa peluang mendapatkan leads di atas 6.000 orang dalam 1x ads?
Leads > 6.000 dikategorikan high

Joint PMF Ads	Leads			
Impression	High	Low	Medium	Grand Total
High	0,125	0,1	0,119	0,299
Low	0,041	0	0,043	0,1
Medium	0,234	0,1	0,237	0,6
Grand Total	0,4	0,2	0,399	1

P(L>6000) = P(L=High) untuk semua nilai impression
P(L=High) = P(L=High, I=High) + P(L=High, I=Low) + P(L=High, I=Medium)
= 0.125 + 0.041 + 0.234
= 0.4

Jadi, peluang mendapatkan leads di atas 6.000 orang dalam 1 kali ads adalah 0.4 atau 40%. Jika ada 100 ads, maka 40 ads berpeluang mendapatkan leads di atas 6.000 orang dalam 1 kali penayangan ads.

d. [3 points] Berapa peluang mendapatkan leads di atas 6.000 orang dalam 1x ads jika diketahui impresi dari ads tersebut kurang dari sama dengan 600K?

Karena diketahui Kondisi = A -> jika diketahui impresi dari ads tersebut kurang dari sama dengan 600K
Maka ruang sample baru hanya Impression Low + Medium
Berikut Joint PMF baru tanpa baris impression high

Count of ads_id	Leads			
Impression	High	Low	Medium	Total
Low	61	24	65	150
Medium	351	193	356	900
Total	412	217	421	1050

Joint PMF | A	Leads			
Impression	High	Low	Medium	Grand Total
Low	0,058	0,023	0,062	0,143
Medium	0,334	0,184	0,339	0,857
Grand Total	0,392	0,207	0,401	1

P(L=High | A) = P(L=High | I = Low dan Medium)
= P(L=High, I = Low) + P(L=High, I=Medium)
= 0.058 + 0.334
= 0.392

Jadi, peluang mendapatkan leads di atas 6.000 orang dalam 1 kali ads jika diketahui impresi dari ads <= 600.000 adalah 0.392 atau 39.2%. Jika ada 100 ads, maka 39 ads berpeluang mendapatkan leads di atas 6.000 jika diketahui impresi dari ads <=600.000

e. [3 points] Sekarang Anda akan membuat rencana marketing sederhana. Diketahui, rata-rata conversion rate dari leads menjadi buyer dalam 1 tahun terakhir adalah 8.5%. Dengan target revenue Rp 300.000.000, Anda harus mendapatkan buyer sebanyak 150 orang. Hitung jumlah leads yang dibutuhkan.

Diketahui Target Revenue = Rp 300.000.000
Conversion rate = 8,5%
Target buyer = 150 orang
Dapat diartikan bahwa dari semua N leads yang didapat, 8.5% akan menjadi buyer. JIka 150 orang adalah 8.5% dari N leads, maka berapa N leads yang dibutuhkan?

Untuk menghitung jumlah leads yang dibutuhkan maka kita perlu membagi jumlah buyer dengan conversion rate

Jumlah leads yang dibutuhkan = Target buyer / conversion rate
= 150 / 8.5%
= 150 / 0.085
= 1765 leads

Jadi, jumlah leads yang dibutuhkan adalah 1765 leads agar dapat mencapai target revenue Rp 300.000.000

f. [3 points] Hitung besar peluang untuk mendapatkan leads tersebut dalam 1x ads.

Menurut kategori leads, Leads sejumlah 1.765 orang tergolong leads dengan kategori Low. Maka, yang akan kita hitung adalah P(L=Low) untuk semua kategori impressions.

Joint PMF Ads	Leads			
Impression	High	Low	Medium	Grand Total
High	0,125	0,055	0,119	0,299
Low	0,041	0,016	0,043	0,1
Medium	0,234	0,129	0,237	0,6
Grand Total	0,4	0,2	0,399	1

P(Leads >=1765) = P(L=Low) untuk semua nilai Impressions

P(Leads>=1765) = 0.055 + 0.016 + 0.129
= 0.2

Jadi, peluang untuk mendapatkan 1765 leads dalam 1x ads adalah sebesar 0.2 atau 20%.

g. [3 points] Untuk mendapatkan target leads dalam 1x ads, Anda dapat menggunakan Ads yang impresinya berkategori medium atau high. Berdasarkan data historis yang Anda punya, Ads dengan kategori apa yang memiliki peluang lebih besar untuk mendapatkan target leads tersebut?

Untuk menentukan kategori impresi maka kita perlu membandingkan P(L=Low | I=Medium) dengan P(L=Low | I=High)

Berikut merupakan jumlah ads dan joint PMF untuk kategori Leads Low di Impresi High dan Medium

Count of ads_id	Leads			
Impression	High	Low	Medium	Grand Total
High	188	83	179	450
Low	61	24	65	150
Medium	351	193	356	900
Grand Total	600	300	600	1500
Joint PMF Ads	Leads			
Impression	High	Low	Medium	Grand Total
High	0,125	0,055	0,119	0,299
Low	0,041	0,016	0,043	0,1
Medium	0,234	0,129	0,237	0,6
Grand Total	0,4	0,2	0,399	1

Dapat dilihat bahwa berdasarkan historis, P(L=Low | I=Medium) bernilai 0.129 dan lebih besar dari P(L=Low | I=High) yang hanya sebesar 0.055. Maka dari itu, **ads dengan impression medium lebih besar peluangnya untuk mendapatkan target leads dengan peluang sebesar 12.9%**

h. [3 points] Apabila harga ads dengan impresi medium adalah Rp 500.000 / ads, dan harga ads dengan impresi high adalah Rp 1.500.000 / ads, berapa ekspektasi biaya ads yang harus dikeluarkan agar mendapatkan target leads tersebut? (Saat menghitung ekspektasi, asumsikan gunakan batas atas jumlah leads pada kategori terkait)

Karena peluang ads medium lebih besar, maka berikut ekspektasi biaya ads dengan impresi medium untuk mendapat target leads >=1765

Diketahui peluang ads medium mendapatkan target leads adalah 0.129 atau hanya 12.9% ads Medium yang berhasil hasilkan ≥1.765 leads
Ekspektasi jumlah ads agar berhasil = 1 / 0.129
≈ 7.75 ads atau dibulatkan menjadi 8 ads
Total biaya ads = 8 × Rp 500.000 = Rp 4.000.000

Jadi, ekspektasi biaya ads yang harus dikeluarkan adalah sebesar Rp 4.000.000

i. [3 points] Terakhir, buat rangkuman rencana marketing sederhana berikut dari jawaban soal-soal di atas.

- Target revenue : Rp 300.000.000
- Conversion rate : 8.5 % (asumsi)
- Target leads : 1765
- Peluang dapat target leads untuk 1x ads : 0.2 atau 20%
- Target impresi ads : Medium Impressions
- Peluang dapat target leads untuk 1x ads dengan impresi target : 12.9%
- Jumlah ads yang dibutuhkan : 8
- Budget marketing yang dibutuhkan : Rp 4.000.000

3. [12 points] Anda bekerja sebagai product analyst di suatu industri fintech. Industri Anda memiliki dua produk unggulan, produk A dan produk B. Saat ini perusahaan Anda diterpa badai akibat lesunya kondisi ekonomi masyarakat. Dampaknya, Anda diminta oleh manager untuk merekomendasikan 1 produk (antara produk A dan B) yang harus dipertahankan. Sebagai tambahan, manager meminta Anda agar produk yang terpilih dapat konsisten menghasilkan revenue.

Untuk menjawab pertanyaan manager, Anda mengumpulkan data conversion rate bulanan, dalam 40 bulan terakhir, antara produk A dan B. Data dapat Anda lihat di-link berikut: .csv . Berikut adalah tampilan dari data yang dimaksud.

	date	Produk A	Produk B
0	2020-01-31 00:00:00	0.504967	0.726405
1	2020-02-29 00:00:00	0.498617	0.590016
2	2020-03-31 00:00:00	0.506477	0.647874
3	2020-04-30 00:00:00	0.51523	0.774089
4	2020-05-31 00:00:00	0.497658	0.736756

Rincian feature:

- date : waktu pencatatan CVR produk. Selalu di akhir bulan.
- Produk A : CVR bulanan dari produk A

- Produk B : CVR bulanan dari produk B

Pertanyaannya

a. [6 points] Produk mana yang Anda pertahankan?

1. Tambahkan dataset ke python

```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd

file_path = '/content/drive//My Drive/PACMANN/PORTOFOLIO/PROBABILITY/cvr_df.csv'
df = pd.read_csv(file_path)

# Lihat 5 baris pertama
df.head()
```

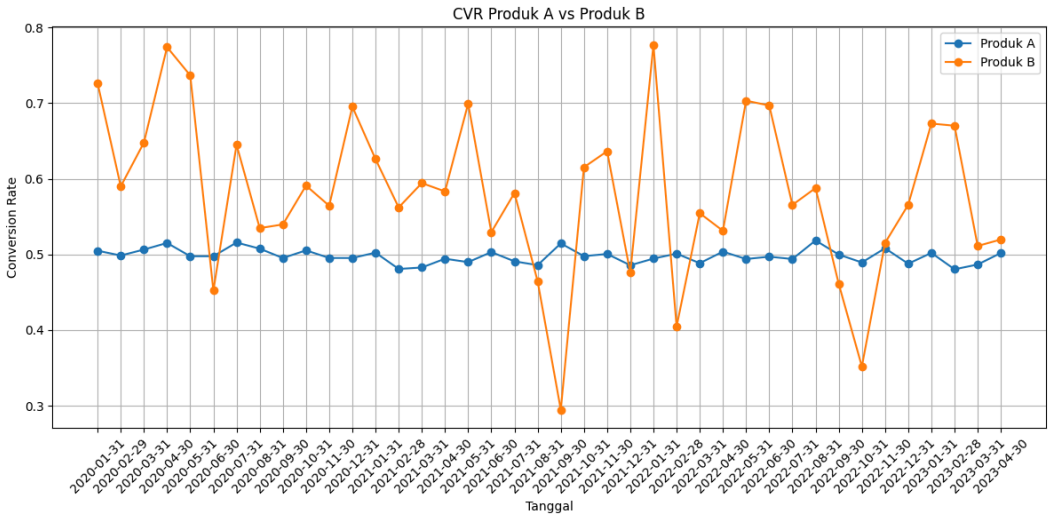
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

	date	Produk A	Produk B
0	2020-01-31	0.504967	0.726405
1	2020-02-29	0.498617	0.590016
2	2020-03-31	0.506477	0.647874
3	2020-04-30	0.515230	0.774089
4	2020-05-31	0.497658	0.736756

2. Plot data untuk melihat pola tren conversion rate antara 2 jenis produk tersebut

```
#Plot time series untuk memahami tren dan fluktuasi
import matplotlib.pyplot as plt

plt.figure(figsize=(12,6))
plt.plot(df['date'], df['Produk A'], label='Produk A', marker='o')
plt.plot(df['date'], df['Produk B'], label='Produk B', marker='o')
plt.xlabel('Tanggal')
plt.ylabel('Conversion Rate')
plt.title('CVR Produk A vs Produk B')
plt.legend()
plt.grid(True)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

3. Hasil Plot



4. Hitung statistik deskriptif dari 2 produk

```
[38] # Hitung statistik deskriptif dari tiap produk

     print(df[['Produk A', 'Produk B']].describe())
```

	Produk A	Produk B
count	40.000000	40.000000
mean	0.497814	0.581254
std	0.009528	0.107789
min	0.480403	0.294701
25%	0.490658	0.527056
50%	0.497659	0.582337
75%	0.503296	0.653465
max	0.518523	0.776975

5. Kesimpulan
   Dari 2 produk tersebut, **lebih baik mempertahankan produk A.**

b. [6 points] Apa alasan Anda mempertahankan produk tersebut?

Alasan mempertahankan produk A:

1. Dapat dilihat dari plot data dan juga statistik deskriptif bahwa produk A memiliki standar deviasi yang lebih kecil yaitu 0.0095 dibanding produk B yang lebih besar yaitu 0.1077. Hal ini menggambarkan bahwa conversion rate produk A lebih konsisten selama 40 bulan.
2. Dapat dilihat dari plot data dan juga statistik deskriptif, bahwa produk B cenderung memiliki conversion rate yang fluktuatif. CVR produk B pernah serendah 0.29 dan setinggi 0.77. Dalam kondisi ketidakstabilan ekonomi, fluktuasi tinggi dapat berisiko tinggi. Hal ini membuat revenue lebih sulit diprediksi.
3. Dengan kondisi ekonomi yang tidak stabil, stabilitas conversion rate lebih penting. Walau conversion rate produk A sedikit lebih rendah dari produk

B, namun produk A lebih dapat diprediksi serta minim risiko sehingga cocok dipertahankan perusahaan dalam situasi ekonomi yang tidak stabil.

4. [24 points] Di suatu perusahaan SaaS AI, tim Sales mendapatkan target revenue bulan ini sebesar Rp 2.000.000.000. Perusahaan tersebut memiliki 2 paket jasa, yaitu Gold Package dan Silver Package. Anda sebagai marketing analyst diminta untuk mencari tau apakah bulan ini tim Sales dapat mengejar targetnya tanpa mencari leads baru atau tidak.

   Anda terpikir untuk gunakan data historis subscription.csv yang berisi informasi waktu user pertama kali jadi leads dan user pertama kali membeli paket jasa untuk melakukan analisa.

user_id	date_leads	date_convert	category
18364	2024-04-01 00:00:38	2024-04-04 03:51:02	Silver Package
13783	2024-04-01 00:01:09	2024-04-05 22:20:21	Silver Package
3112	2024-04-01 00:01:17	2024-04-01 17:46:53	Gold Package
21076	2024-04-01 00:01:24	2024-04-02 05:32:36	Silver Package
12005	2024-04-01 00:01:56	2024-04-09 15:23:32	Gold Package

   Feature:
   - user_id : ID user
   - date_leads : Waktu pertama user menjadi leads
   - date_convert : Waktu pertama user convert (membeli produk)
   - category : Kategori paket jasa yang dibeli

   Selesaikan rangkaian pertanyaan berikut untuk menjawab pertanyaan:
   a. [4 points] Kita bisa mencari waktu yang dibutuhkan untuk user berubah dari leads menjadi convert (days_to_convert). Olah data yang diberikan dan hitung informasi days_to_convert untuk masing-masing user. Ini akan menjadi random variabel yang akan kita olah selanjutnya. **Biarkan dalam bentuk float agar random variable nya berbentuk kontinu**.
      1. Tambah dataset ke python

```
[3] from google.colab import drive
    drive.mount('/content/drive')

    import pandas as pd

    file_path = '/content/drive//My Drive/PACMANN/PORTOFOLIO/PROBABILITY/subscription.csv'
    df = pd.read_csv(file_path)

    # Lihat 5 baris pertama
    df.head()
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

	user_id	date_leads	date_convert	category
0	18364	2024-04-01 00:00:38	2024-04-04 03:51:02	Silver Package
1	13783	2024-04-01 00:01:09	2024-04-05 22:20:21	Silver Package
2	3112	2024-04-01 00:01:17	2024-04-01 17:46:53	Gold Package
3	21076	2024-04-01 00:01:24	2024-04-02 05:32:36	Silver Package
4	12005	2024-04-01 00:01:56	2024-04-09 15:23:32	Gold Package

2. Tambahkan data days_to_convert ke dataset

```
[31] #Tambah kolom days_to_convert ke dataset

     #Memastikan data_leads dan data_convert dibaca sebagai datetime
     df['date_leads'] = pd.to_datetime(df['date_leads'])
     df['date_convert'] = pd.to_datetime(df['date_convert'])

     #Menghitung days_to_convert
     df['days_to_convert'] = (df['date_convert'] - df['date_leads']).dt.days
     df['days_to_convert'] = (df['date_convert'] - df['date_leads']).dt.total_seconds() / (3600 * 24)
     df[['user_id', 'date_leads', 'date_convert', 'days_to_convert']].head()

     # Simpan ke file baru
     output_path = '/content/drive/My Drive/PACMANN/PORTOFOLIO/PROBABILITY/subscription_with_days.csv'
     df.to_csv(output_path, index=False)

     # Load ulang file yang sudah ditambah kolom
     df = pd.read_csv(output_path)

     # Pastikan kolom tanggal tetap diubah ke datetime
     df['date_leads'] = pd.to_datetime(df['date_leads'])
     df['date_convert'] = pd.to_datetime(df['date_convert'])

     # Lihat 5 baris pertama
     df.head()
```

	user_id	date_leads	date_convert	category	days_to_convert
0	18364	2024-04-01 00:00:38	2024-04-04 03:51:02	Silver Package	3.16
1	13783	2024-04-01 00:01:09	2024-04-05 22:20:21	Silver Package	4.93
2	3112	2024-04-01 00:01:17	2024-04-01 17:46:53	Gold Package	0.74
3	21076	2024-04-01 00:01:24	2024-04-02 05:32:36	Silver Package	1.23
4	12005	2024-04-01 00:01:56	2024-04-09 15:23:32	Gold Package	8.64

b. [4 points] Kita akan fokus menganalisa konversi masing-masing kategori paket secara terpisah. Cari tau tipe distribusi random variable days_to_convert untuk masing-masing kategori!

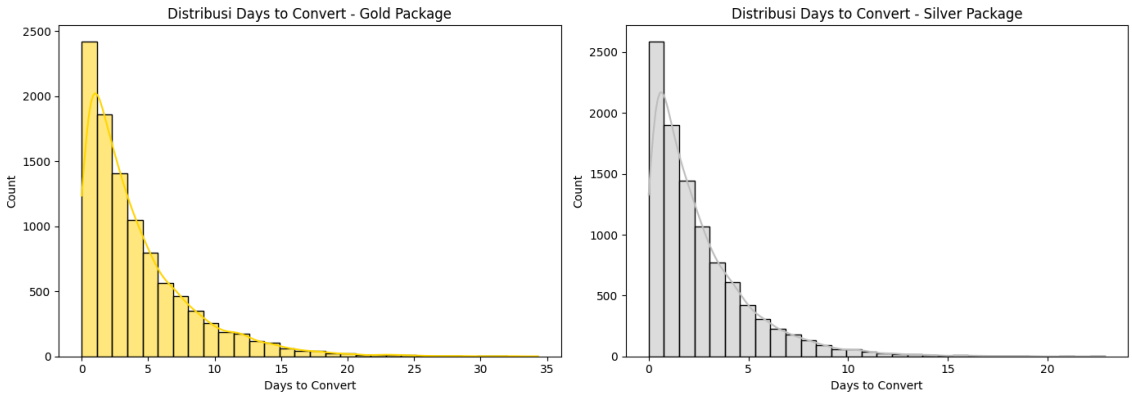
1. Filter data berdasarkan kategori package

```
[29] # Filter kategori

     gold = df[df['category'] == 'Gold Package']
     silver = df[df['category'] == 'Silver Package']
```

2. Gambarkan plot distribusi

```
[30] import matplotlib.pyplot as plt
     import seaborn as sns

     plt.figure(figsize=(14, 5))

     # Gold
     plt.subplot(1, 2, 1)
     sns.histplot(gold['days_to_convert'], kde=True, bins=30, color='gold')
     plt.title('Distribusi Days to Convert - Gold Package')
     plt.xlabel('Days to Convert')

     # Silver
     plt.subplot(1, 2, 2)
     sns.histplot(silver['days_to_convert'], kde=True, bins=30, color='silver')
     plt.title('Distribusi Days to Convert - Silver Package')
     plt.xlabel('Days to Convert')

     plt.tight_layout()
     plt.show()
```

3. Hasil plot distribusi



4. Tentukan tipe distribusi random variabel days_to_convert

   Dari bentuk histogram, kita bisa simpulkan **distribusi keduanya cenderung exponential** karena positif-skewed dan turun tajam dari hari ke-0. Ini menunjukkan: **semakin cepat hari ke hari**, semakin kecil kemungkinan user untuk convert. Jadi, banyak user convert di hari-hari awal sejak jadi leads.

c. [4 points] Carilah berapa rata-rata days_to_convert dari masing-masing kategori paket

   Rata-data Days to convert
   Gold = 4,105
   Silver = 2,533

```
# Hitung rata-rata
print("Rata-rata Days to Convert:")

#Mengelompokkan baris di  dataset berdasarkan category dengan groupby
# dihitung "mean" nya
print(df.groupby('category')['days_to_convert'].mean())
```

```
Rata-rata Days to Convert:
category
Gold Package      4.105486
Silver Package    2.533669
Name: days_to_convert, dtype: float64
```

d. [8 points] Diketahui ada 1.500 leads yang telah menjadi leads selama 5 hari. Tentukan pilihan paket yang direkomendasikan untuk 1.500 leads tersebut. *Hint: Cari probability leads convert*

Berdasarkan analisis sebelumnya, distribusi **days_to_convert** adalah **eksponensial**, sehingga probabilitas konversinya dihitung menggunakan **CDF distribusi eksponensial** yaitu :

$$P(X \leq x) = 1 - e^{-\lambda x}$$

λ = 1 / rata-rata days_to_convert
x = 5 hari

Gold Package

Rata-rata days_to_convert	= 4,105
λ_{gold}	= 1 / 4,105
	= 0,244
CDF pada x	= 5
P_{gold}	= $1 - e^{-0.244x5}$
	= $1 - e^{-1.22}$
	= 1 - 0,295
	= 0,705

Jadi, probabilitas convert gold package adalah 70,5%

Silver Package

Rata-rata days_to_convert	= 2,533
λ_{gold}	= 1 / 2,533
	= 0.395
CDF pada x	= 5
P_{gold}	= $1 - e^{-0.395x5}$

$= 1 - e^{-1,975}$
$= 1 - 0,139$
$= 0,861$

Jadi, probabilitas convert silver package adalah 86,1%

Berdasarkan perbandingan probabilitas konversi kedua paket maka dipilih silver package karena probabilitasnya lebih besar.

e. [2 points] Apabila harga subscription untuk paket Gold adalah Rp 3.750.000 / user dan paket Silver adalah Rp 1.500.000 / user. Tentukan berapa ekspektasi revenue apabila user di soal sebelumnya convert.

Expected Revenue = P(Silver) x Harga x Jumlah leads
= 0.861 x 1.500.000 x 1.500
= Rp1.937.250.000

Jadi, Ekspektasi revenue bila memilih Silver Package adalah Rp 1.937.250.000.

f. [2 points] Dari jawaban soal sebelumnya, apakah target tim Sales bulan ini dapat tercapai tanpa mencari leads baru? Elaborasikan jawabannya.

Target tim Sales bulan ini sebesar Rp 2.000.000.000 tidak akan tercapai sepenuhnya tanpa mencari leads baru.

Berdasarkan perhitungan, ekspektasi revenue dari 1.500 leads yang telah menjadi leads selama 5 hari jika ditawarkan Silver Package hanya mencapai Rp 1.937.250.000, sehingga masih terdapat kekurangan sebesar Rp 62.750.000.

Artinya, meskipun peluang konversinya tinggi, tim Sales masih memerlukan tambahan leads atau strategi lain (seperti upselling atau mempercepat konversi leads yang lebih lama) untuk benar-benar mencapai target.

Namun demikian, bila tim Sales mempertimbangkan untuk menawarkan Gold Package, memang secara nominal revenue yang dihasilkan lebih besar karena harga paket yang lebih tinggi. Dengan probabilitas konversi sekitar 70,35% dan harga Gold Package yang lebih mahal, estimasi revenue bisa mencapai Rp 3.955.312.500 dari 1.500 leads yang sama. Namun, ada beberapa hal yang perlu diperhatikan. Memilih Gold Package berisiko menghasilkan lebih sedikit pelanggan yang berhasil convert karena peluang konversinya lebih rendah, serta bisa jadi tidak semua leads memiliki daya beli atau kebutuhan yang sesuai untuk paket Gold. Selain itu, Silver Package memiliki peluang konversi yang lebih tinggi (86,10%), sehingga lebih realistis untuk dimaksimalkan dalam jangka pendek tanpa mengorbankan kepuasan dan kepercayaan leads.

Dengan mempertimbangkan risiko dan prioritas keberhasilan konversi, maka paket yang direkomendasikan tetap **Silver Package**. Meskipun target revenue bulan ini belum sepenuhnya tercapai, pendekatan ini lebih aman dan berpotensi memperbesar basis pelanggan baru. Untuk menutupi kekurangan revenue, tim Sales dapat melengkapi strategi ini dengan upaya tambahan seperti upselling ke pelanggan existing atau mempercepat konversi leads lain yang belum ditargetkan secara maksimal.

5. [27 points] Anda sedang melakukan analisa customer di industri ecommerce. Anda ingin memodelkan peluang customer melakukan pembelian lagi 2 hari setelah pembelian terakhir sebelum user tersebut churn. Untuk itu, Anda membuat model peluang customer tersebut dengan mendefinisikan 2 random variable dengan rincian
   - Model 1 - mendefinisikan user tidak aktif tepat setelah transaksi ke - x
   - Model 2 - mendefinisikan waktu user aktif melakukan pembelian dalam waktu t setelah pembelian terakhir.

   Pertanyaannya:
   - [10 points] Tentukan distribusi variabel random yang dapat memodelkan 2 kondisi di atas! Sertakan alasannya.

     Model 1 : User tidak aktif tepat setelah transaksi ke - x
     Distribusi yang cocok : Distribusi Geometrik
     - Memodelkan jumlah trial (dalam hal ini transaksi) sampai terjadinya kegagalan pertama (churn).
     - Cocok dengan sifat distribusi geometrik yang memodelkan jumlah percobaan hingga kegagalan pertama.
     - Probabilitas churn setelah setiap transaksi konstan.
     - Setiap trankasi diasumsikan memiliki peluang tetap P agar user tidak melakukan transaksi lagi setelehnya.

     Model 2 : Waktu user aktif melakukan pembelian dalam waktu t setelah pembelian terakhir
     Distribusi yang cocok : Distribusi Eksponensial
     - Memodelkan waktu antar transaksi
     - Cocok dengan distribusi eksponensial yang memodelkan waktu tunggu hingga terjadinya suatu event berikutnya. Dalam hal ini, memodelkan waktu antara dua transaksi oleh pelanggan.
     - Jika proses pembelian user mengikuti pola poisson dengan pembelian sebagai event, maka waktu antar pembelian mengikuti distribusi eksponensial

   - [17 points] Buat model untuk mencari peluang
     - User melakukan transaksi ke-**x**,

- Kurang dari sama dengan $\mathbf{t_1}$ hari setelah transaksi ke-**x-1**
- Sebelum user menjadi tidak aktif *(Tambahan: user menjadi tidak aktif saat transaksi x selesai dilakukan)*
- Jika diketahui transaksi ke-**x-1** membutuhkan waktu kurang dari sama dengan $\mathbf{t_2}$ setelah transaksi ke-**x-2**.
- Asumsi $t_2 > t_1$

Misalkan
T, ~ Eksponensial (λ) adalah waktu antar transaksi ke-i dan ke-i + 1
X ~ Geometrik(p) adalah jumlah transaksi sampai churn

Maka kita ingin melakukan pemodelan
$P(T_{x-1,x} \leq t_1 \mid T_{x-2,x-1} \leq t_2)$. P(churn setelah transaksi ke-x)

1. Peluang waktu antar transaksi ke-x dan $x-1 \leq t_1$:

   Karena T~ Eksponensial (λ), maka:
   $P(T_{x-1,x} \leq t_1) = 1 - e^{-\lambda t1}$

2. Peluang waktu antar transaksi x–2 dan $x-1 \leq t_2$:
   $P(T_{x-2,x} \leq t_1) = 1 - e^{-\lambda t2}$

3. Kita ingin peluang bersyarat:
   $P(T_{x-1,x} \leq t_1 \mid T_{x-2,x-1} \leq t_2)$

   Jika asumsi independence antara waktu antar transaksi benar (dalam exponential / poisson process):
   $P(T_{x-1,x} \leq t_1 \mid T_{x-2,x-1} \leq t_2) = P(T_{x-1,x} \leq t_1) = 1 - e^{-\lambda t1}$

4. Peluang user churn setelah transaksi ke-x:

   Karena X ~ Geometrik(p), maka peluang user churn di transaksi ke-x (artinya total transaksi = x) adalah:
   $P(X=x) = (1-p)^{x-1}p$

Gabungkan seluruh model:

$P(\text{transaksi ke-x dalam} \leq t_1 \text{ hari, lalu churn} \mid T_{x-2,x-1} \leq t_2) = 1 - e^{-\lambda t1} \cdot (1-p)^{x-1}p$

Kesimpulan:

- Distribusi geometrik digunakan untuk memodelkan jumlah transaksi sampai user churn.

- Distribusi eksponensial digunakan untuk memodelkan waktu antar transaksi.

- Model gabungan di atas memberikan peluang terjadinya pembelian ke-x dalam $\leq t_1$ hari setelah x–1, dengan syarat user telah melakukan transaksi x–1 dalam $\leq t_2$ hari setelah x–2 dan akan churn setelah transaksi ke-x.