Rakamin Academy

MANDALIKA

Dokumen Laporan
Final Project Stage 1

Anggota :

- Ardilla Safitri
- Jonse Kennedy
- Fakhry Husnul
- Lise Listianti
- Riska Diah N
- Arni Cici Suryani

LATAR BELAKANG MASALAH

- Dalam sebulan terakhir, sebuah perusahaan e-commerce memiliki 16,8% customer yang churn.
- Bagaimana cara tim data scientist untuk membantu perusahaan dalam memprediksi customer yang berpotensi churn?
- Menggunakan analytical approach: membentuk sebuah model yang dapat mendeteksi pelanggan yang berpotensi churn.
- Business Metrics yang digunakan ialah **Churn Rate**.

DESCRIPTIVE ANALYSIS

Rakamin Academy

Informasi General Tentang Dataframe

```
# informasi general tentang dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5630 entries, 0 to 5629
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   CustomerID                   5630 non-null   int64
 1   Churn                        5630 non-null   int64
 2   Tenure                       5366 non-null   float64
 3   PreferredLoginDevice         5630 non-null   object
 4   CityTier                     5630 non-null   int64
 5   WarehouseToHome              5379 non-null   float64
 6   PreferredPaymentMode         5630 non-null   object
 7   Gender                       5630 non-null   object
 8   HourSpendOnApp               5375 non-null   float64
 9   NumberOfDeviceRegistered     5630 non-null   int64
 10  PreferedOrderCat             5630 non-null   object
 11  SatisfactionScore            5630 non-null   int64
 12  MaritalStatus                5630 non-null   object
 13  NumberOfAddress              5630 non-null   int64
 14  Complain                     5630 non-null   int64
 15  OrderAmountHikeFromlastYear  5365 non-null   float64
 16  CouponUsed                   5374 non-null   float64
 17  OrderCount                   5372 non-null   float64
 18  DaySinceLastOrder            5323 non-null   float64
 19  CashbackAmount               5630 non-null   float64
dtypes: float64(8), int64(7), object(5)
memory usage: 879.8+ KB
```

Pengamatan:

1. Data terdiri dari 5630 baris
2. Tidak terdapat tipe data dan nama kolom yang kurang sesuai
3. Tampak beberapa kolom masih memiliki null/missing values (Non-Null Count < jumlah baris) diantaranya Tenure, WarehouseToHome, HourSpendOnApp, OrderAmountHikeFromLastYear, CouponUsed, OrderCount, DaySinceLastOrder

DESCRIPTIVE ANALYSIS

Rakamin Academy

Statistical Summary

```
# pengelompokan kolom berdasarkan jenisnya
nums = ['Churn', 'Tenure', 'CityTier', 'WarehouseToHome', 'HourSpendOnApp', 'NumberOfDeviceRegistered',
        'SatisfactionScore', 'NumberOfAddress', 'Complain', 'OrderAmountHikeFromlastYear', 'CouponUsed',
        'OrderCount', 'DaySinceLastOrder', 'CashbackAmount']
cats = ['PreferredLoginDevice', 'PreferredPaymentMode', 'Gender', 'PreferedOrderCat', 'MaritalStatus']
```

```
# ringkasan statistik dari kolom numerik
df[nums].describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.00	0.0000	1.00
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.00	16.0000	61.00
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.00	3.0000	3.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.00	20.0000	127.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.00	3.0000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.00	4.0000	6.00
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.00	4.0000	5.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.00	6.0000	22.00
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.00	1.0000	1.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.00	18.0000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.00	2.0000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.00	3.0000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.00	7.0000	46.00
CashbackAmount	5630.0	177.223030	49.207036	0.0	145.77	163.28	196.3925	324.99

Pengamatan:

1. Kolom Churn, Tenure, Warehouse To Home, Hour Spend On App, Number Of Device Registered, Satisfaction Score, dan Order Amount Hike From last Year sudah mendekati distribusi normal (median tidak jauh berbeda dengan mean)
2. Kolom City Tier, Number Of Address, Coupon Used, Order Count, Day Since Last Order, dan Cashback Amount tampak skew ke kanan (long-right tail).
3. Kolom Churn dan Complain ternyata bernilai boolean/binary
4. Tidak ada kolom yang bernilai negatif

DESCRIPTIVE ANALYSIS

Rakamin Academy

Categorical Columns

```
# categorical columns
df[cats].describe()
```

	PreferredLoginDevice	PreferredPaymentMode	Gender	PreferedOrderCat	MaritalStatus
count	5630	5630	5630	5630	5630
unique	3	7	2	6	3
top	Mobile Phone	Debit Card	Male	Laptop & Accessory	Married
freq	2765	2314	3384	2050	2986

Pengamatan:

1. Data dinominasi (proporsi lebih dari 50% dari jumlah baris data) oleh penggunaan:
   - Mobile Phone (PreferredLoginDevice)
   - Male (Gender)
   - Married (MaritalStatus)
2. Kolom PreferredPaymentMode dan PreferedOrderCat memiliki kardinalitas (jumlah unique values) yang lebih tinggi (6-7) dibandingkan kolom lainnya.

DESCRIPTIVE ANALYSIS

Value Counting

```
for col in cats:
    print(f'''Value count kolom {col}:''')
    print(df[col].value_counts())
    print()
```

```
Value count kolom PreferredLoginDevice:
Mobile Phone    2765
Computer        1634
Phone           1231
Name: PreferredLoginDevice, dtype: int64

Value count kolom PreferredPaymentMode:
Debit Card          2314
Credit Card         1501
E wallet             614
UPI                  414
COD                  365
CC                   273
Cash on Delivery     149
Name: PreferredPaymentMode, dtype: int64

Value count kolom Gender:
Male      3384
Female    2246
Name: Gender, dtype: int64

Value count kolom PreferedOrderCat:
Laptop & Accessory    2050
Mobile Phone          1271
Fashion                826
Mobile                 809
Grocery                410
Others                 264
Name: PreferedOrderCat, dtype: int64

Value count kolom MaritalStatus:
Married     2986
Single      1796
Divorced     848
Name: MaritalStatus, dtype: int64
```

Pengamatan:

1. Value count kolom Preferred Log in Device: jumlah Phone sama dengan Mobile Phone sehingga datanya harus digabung
2. Value count kolom Preferred Payment Mode: Jumlah Credit Card sama dengan CC dan COD sama dengan Cash on Delivery sehingga datanya harus digabung
3. Value count kolom Preferred Order Cat: Jumlah Mobile Phone sama dengan Mobile

UNIVARIATE ANALYSIS

Setelah melakukan analisis sederhana tentang statistik deskriptif, sekarang kita fokus pada satu-persatu kolom dengan *Univariate Analysis.* Pada Univariate Analisys digunakan Box Plot, Distribution Plot dan Count Plot.

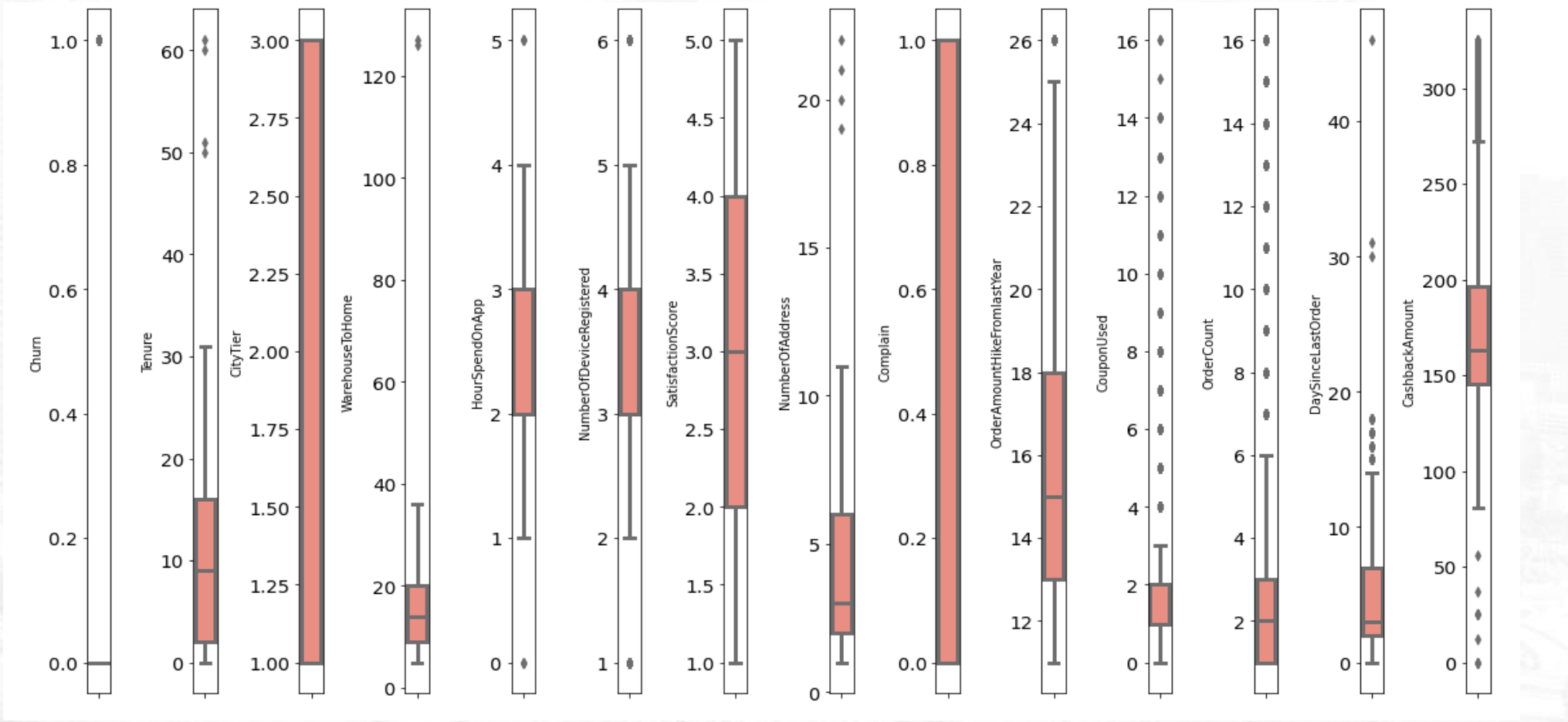
Box Plot

```
features = nums
for i in range(0, len(features)):
    plt.subplot(1, len(features), i+1)
    sns.boxplot(y=df[features[i]], color='salmon', orient='v')
    plt.tight_layout()
```

UNIVARIATE ANALYSIS

Rakamin Academy

Box Plot



Untuk boxplot, hal paling penting yang harus kita perhatikan adalah keberadaan outlier.

1. Outlier terlihat utamanya pada kolom CouponUsed, OrderCount, CashbackAmount dan DaySinceLastOrder
2. Dari boxplotnya juga tampak mana distribusi yang terlihat agak *skewed*: Tenure, WarehouseToHome, NumberOfAddress, CouponUsed, OrderCount, dan DaySinceLastOrder

UNIVARIATE ANALYSIS

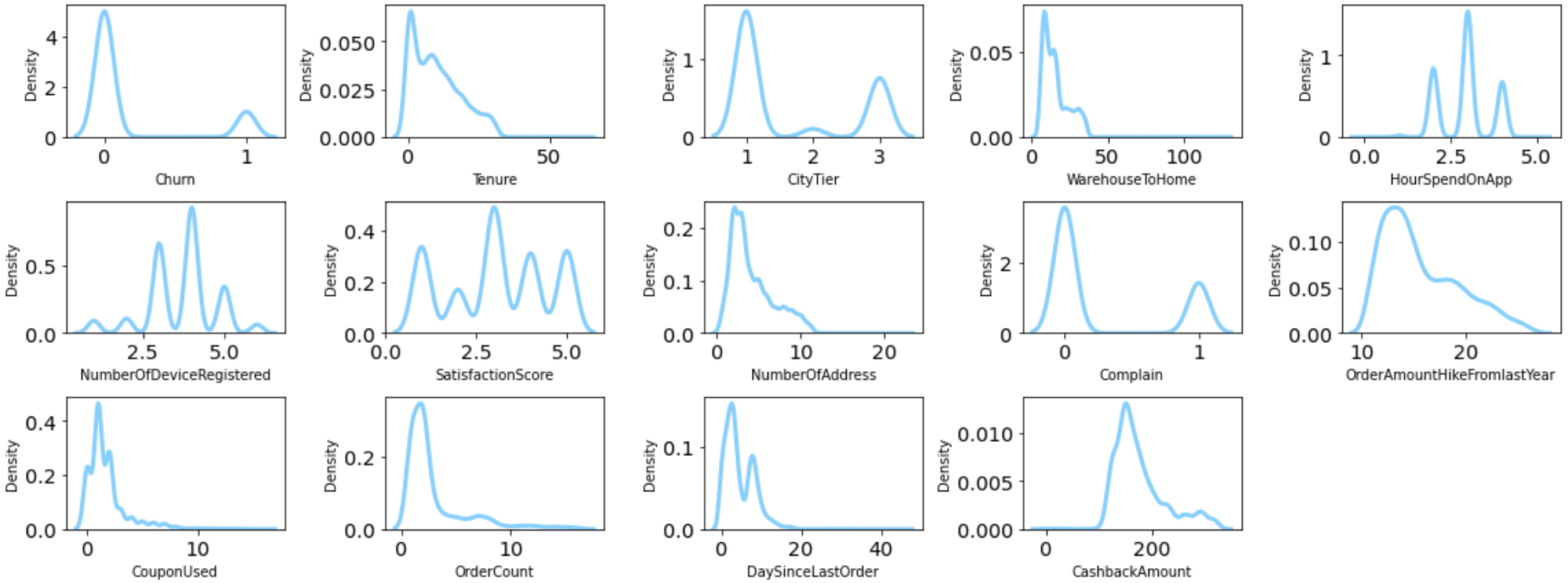
Setelah dilakukan Box Plot untuk melihat Outlier pada setiap kolom data, maka setelah itu dilakukan Distribution Plot untuk mengetahui distribusi dari masing – masing data

Distribution Plot

```
features = nums
plt.figure(figsize=(16, 6))
for i in range(0, len(nums)):
    plt.subplot(3, 5, i+1)
    sns.kdeplot(x=df[features[i]], color='lightskyblue')
    plt.xlabel(features[i])
    plt.tight_layout()
```

UNIVARIATE ANALYSIS

Dist Plot



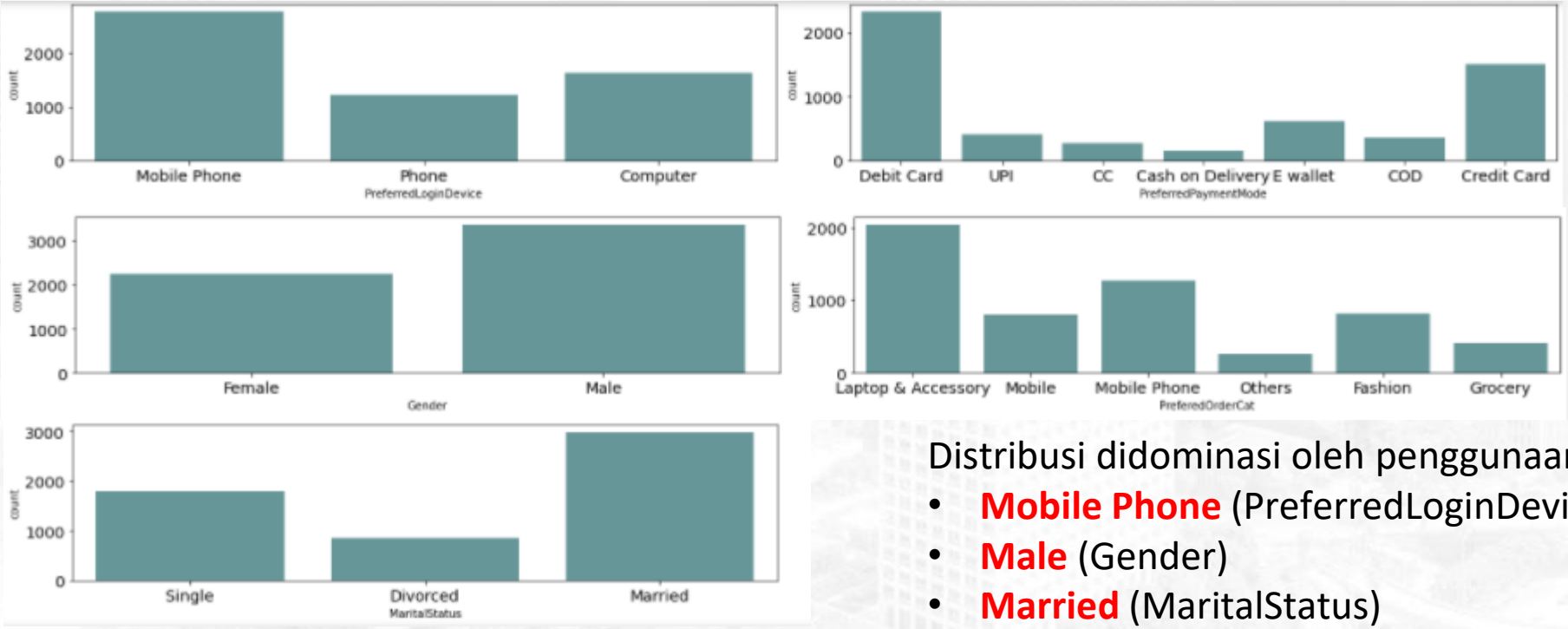
Untuk distribution plot, hal utama yang perlu diperhatikan adalah bentuk distribusi:

- Kolom HourSpendApp, NumberOfDeviceRegistered, SatisfactionScore, dan OrderAmountHikeFromlastYear tampak sudah mendekati distribusi normal
- Seperti dugaan sebelumnya ketika melihat boxplot di atas,
  kolom Tenure, WarehouseToHome, NumberOfAddress, OrderAmountHikeFromlastYear, CouponUsed, OrderCount, dan DaySinceLastOrder sedikit *skewed*
- Berarti ada kemungkinan kita perlu melakukan sesuatu pada kolom-kolom tersebut nantinya
- Kolom churn dan Complain merupakan boolean, sehingga tidak perlu terlalu diperhatikan bentuk distribusinya

UNIVARIATE ANALYSIS

Count plot (categorical)

```
[13]: plt.figure(figsize=(20, 8))
      for i in range(0, len(cats)):
          plt.subplot(3, 2, i+1)
          sns.countplot(x = df[cats[i]], color='cadetblue', orient='v')
          plt.tight_layout()
```



Distribusi didominasi oleh penggunaan:

- **Mobile Phone** (PreferredLoginDevice)
- **Male** (Gender)
- **Married** (MaritalStatus)

*(proporsi lebih dari 50% dari jumlah baris data)

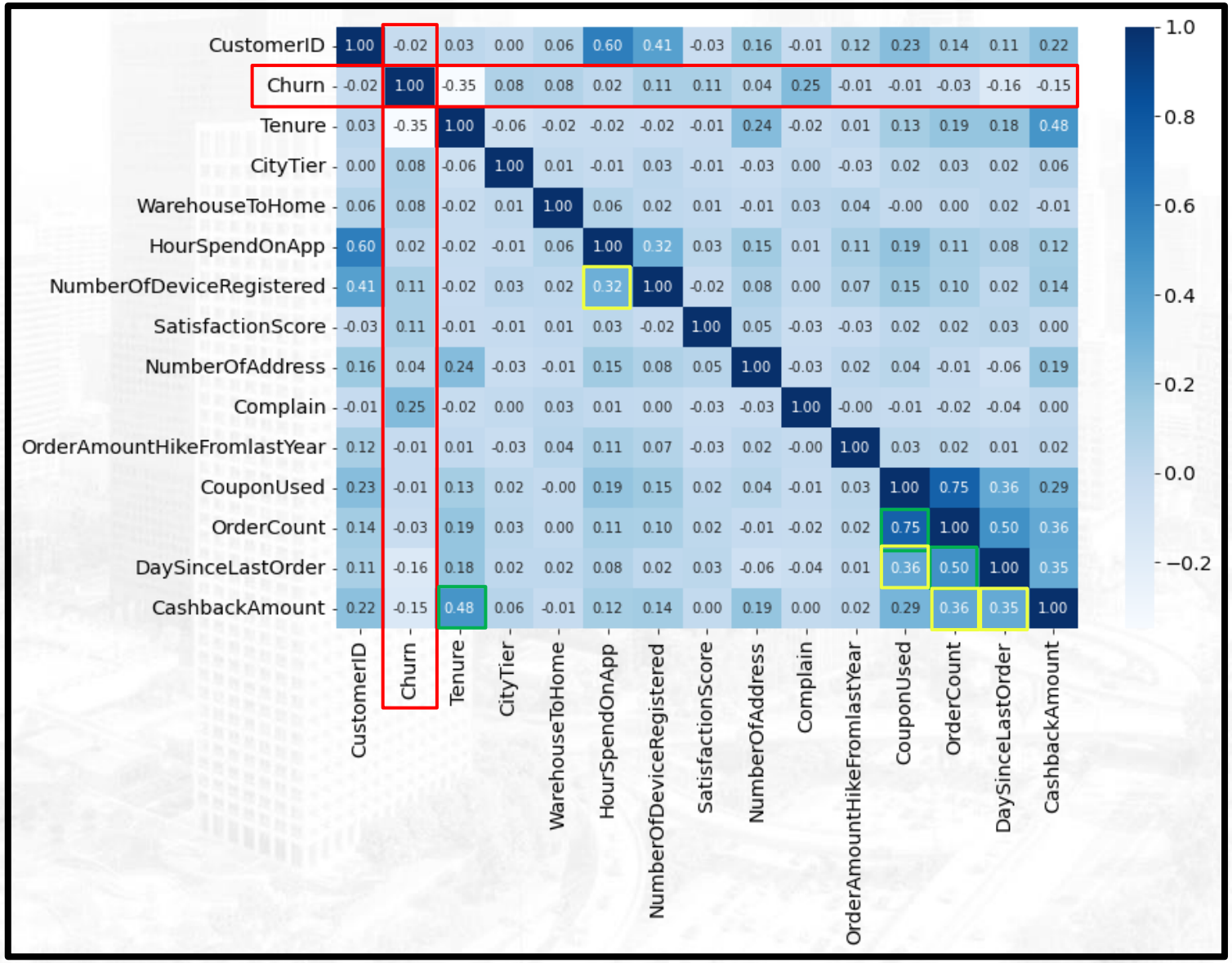
MULTIVARIATE ANALYSIS

Rakamin Academy

Dari correlation heatmap di samping dapat dilihat bahwa:

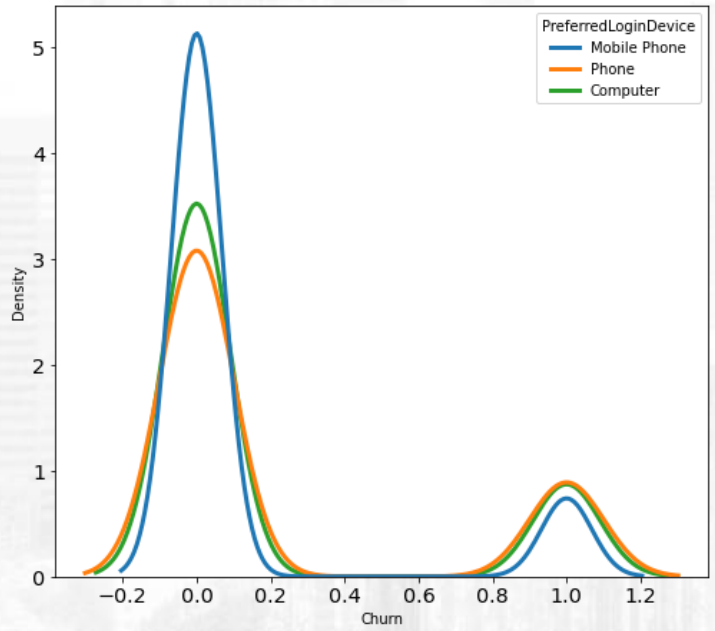
- Target (churn) memiliki korelasi positif lemah dengan Complain, SatisfactionScore, dan NumberDeviceRegistered.
- Target (churn) memiliki korelasi negatif lemah dengan CashbackAmount, DaySinceLater, dan Tenure
- OrderCount dengan CouponUsed (0.75), Tenure dengan CashbackAmount (0.48), DaySinceLastOrder dengan OrderCount (0.5) memiliki korelasi positif kuat. Ada kemungkinan kedua feature ini redundan, sehingga dapat dipilih salah satu saja (dan dibuang yang lainnya).
- NumberOfDeviceRegistered dengan HourSpendOnApp (0.32), CashbackAmount dengan OrderCount (0.36), DaySinceLastOrder dengan CashbackAmount (0.35), dan DaySinceLastOrder dengan CouponUsed (0.36) memiliki korelasi positif lemah.

Correlation Heatmap

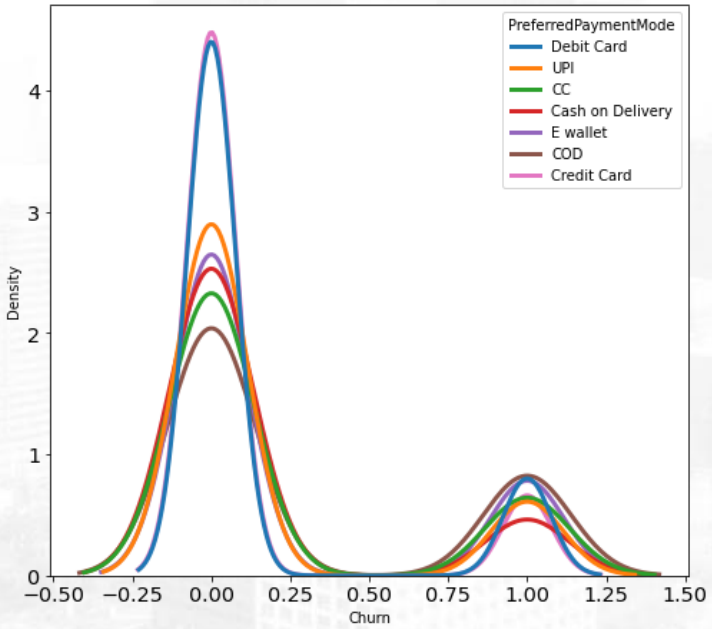


MULTIVARIATE ANALYSIS

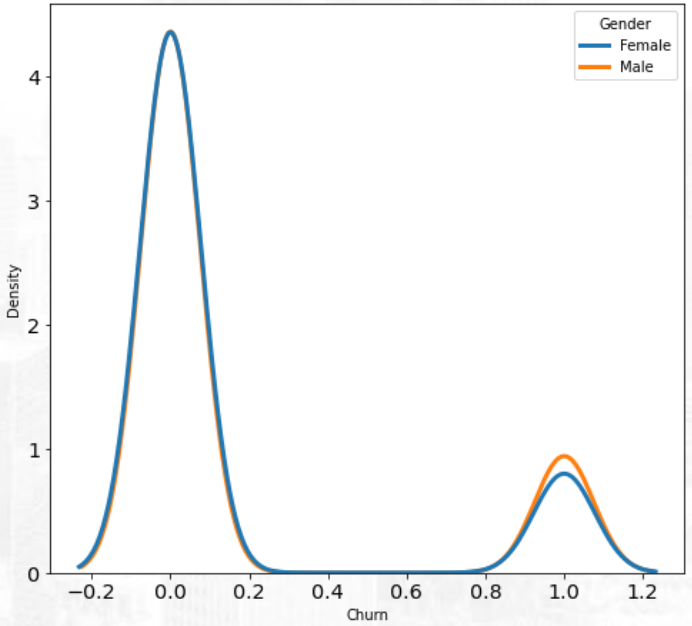
Category Columns VS Target Variable



Pengguna Mobile Phone cenderung sedikit lebih beresiko untuk churn.



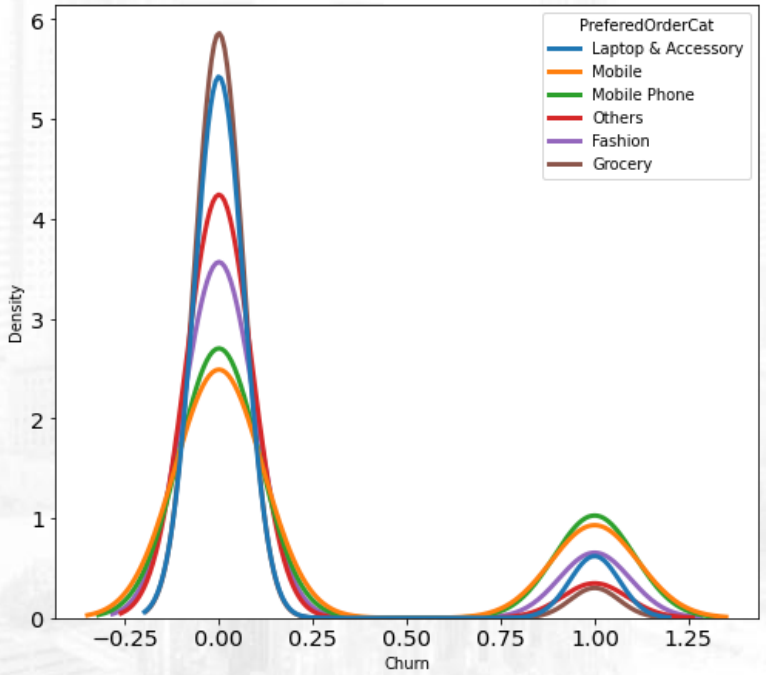
Pembayaran melalui Debit Card cenderung sedikit lebih beresiko untuk churn.



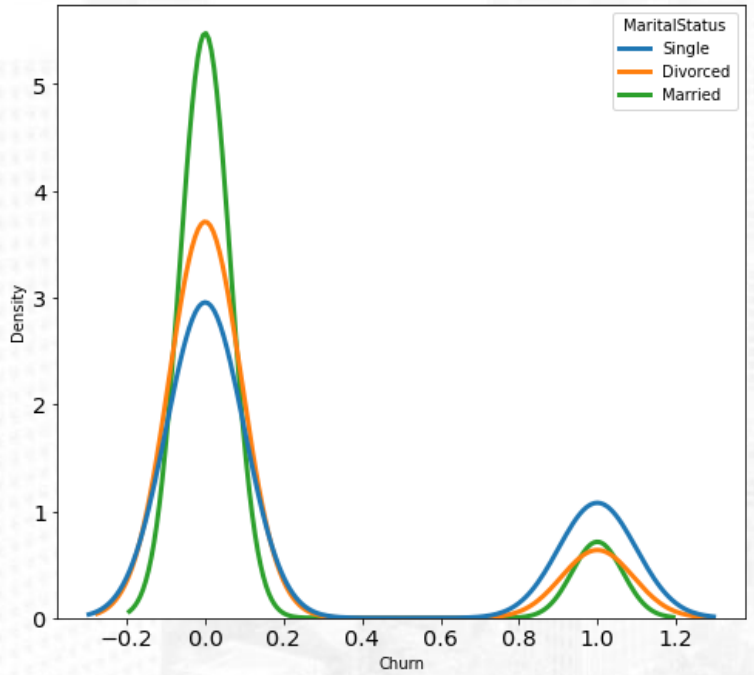
Laki-laki tampaknya cenderung sedikit lebih berisiko untuk churn.

MULTIVARIATE ANALYSIS

Category Columns VS Target Variable



Berdasarkan PreferedOrderCat, Laptop & accessory dan Mobile Phone cenderung sedikit lebih beresiko untuk churn.



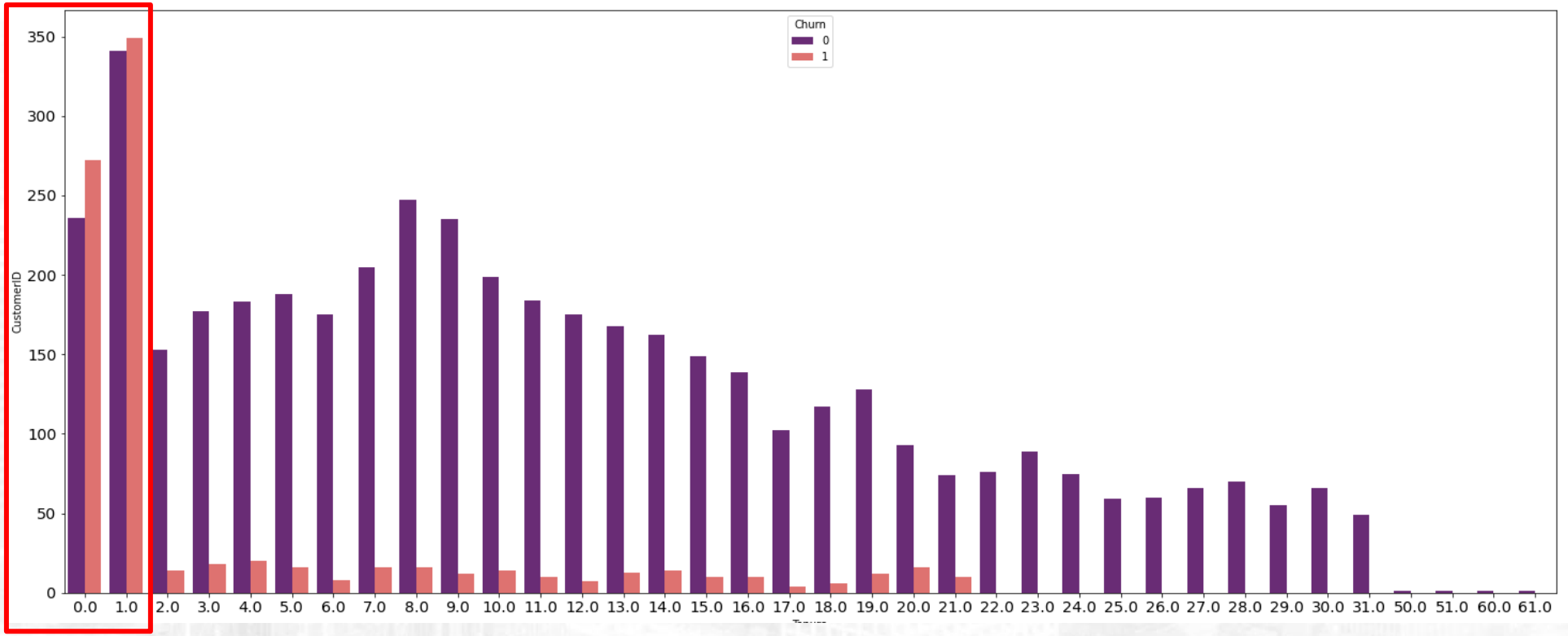
Berdasarkan MaritalStatus, Single cenderung sedikit lebih beresiko untuk churn.

EDA CONCLUSION

Beberapa hal yang ditemukan dari EDA dataset ialah:

- Data terlihat valid dan tidak ada kecacatan yang major/signifikan
- Namun masih ada sedikit data-data yang kosong/hilang & tidak sesuai (Mobile Phone = Phone, Credit Card = CC, dan COD = Cash on Delivery, Mobile Phone=Mobile) Harus diurus ketika preprocessing
- Ada beberapa distribusi yang sedikit *skewed*, hal ini harus diingat apabila kita ingin melakukan sesuatu atau menggunakan model yang memerlukan asumsi distribusi normal
- Beberapa *feature* memiliki korelasi lemah dengan *target*, sehingga mereka akan dipakai (Complain, SatisfactionScore, NumberDeviceRegistered, CashbackAmount, DaySinceLastOrder, dan Tenure)
- Beberapa *feature* terlihat sama sekali tidak berkorelasi, mereka sebaiknya diabaikan (OrderCount & CouponUsed, Tenure & CashbackAmount, DaySinceLastOrder, dan OrderCount)

BUSINESS INSIGHT



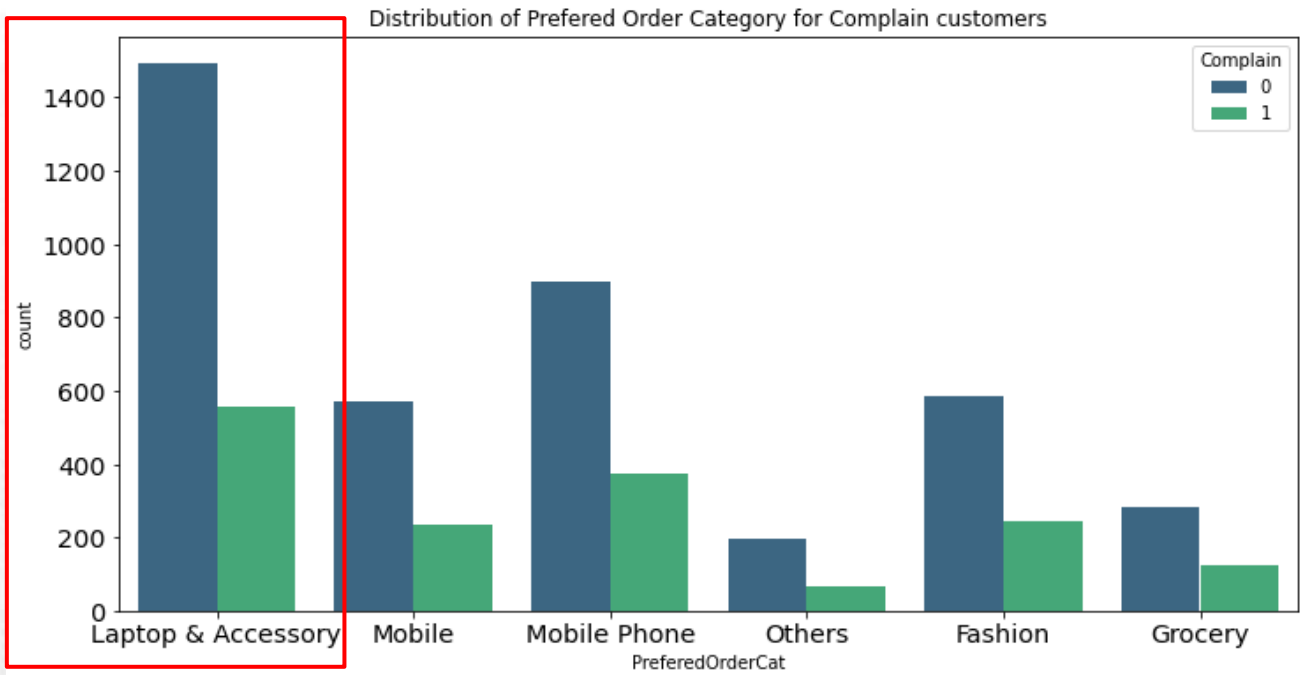
Kesimpulan:

Dilihat dari feature Tenure, pengguna baru yaitu pengguna yang memiliki Tenure < 1 minggu lebih beresiko untuk churn

BUSINESS INSIGHT

```
dfp = df.groupby(['Gender', 'MaritalStatus', 'Churn'])['CustomerID'].count().reset_index()
dfp
```

	Gender	MaritalStatus	Churn	CustomerID
0	Female	Divorced	0	300
1	Female	Divorced	1	48
2	Female	Married	0	1028
3	Female	Married	1	112
4	Female	Single	0	570
5	Female	Single	1	188
6	Male	Divorced	0	424
7	Male	Divorced	1	76
8	Male	Married	0	1614
9	Male	Married	1	232
10	Male	Single	0	746
11	Male	Single	1	292



Kesimpulan:

Mayoritas customer churn berdasarkan gender dan marital status ialah laki-laki single sehingga perlu adanya spesial service untuk customer laki-laki yang masih single.

Kesimpulan:

Berdasarkan Prefered Order Category untuk kategori Laptop & Accessory memiliki complain yang lebih banyak dibandingkan kategori yang lain, sehingga beresiko untuk churn.

CONCLUSION

1. Apa saja attributes dan target output dari dataset yang dipilih?

Attributes: Complain, Satisfaction Score, Cashback Amount, Day Since Last Order, dan Tenure

Target : Churn

2. Untuk setiap feature yang disiapkan, apakah sudah dicek distribusinya terhadap variabel target?

Complain : Tipe data boelean

SatisfactionScore : Normal

CashbackAmount : Positive skewed

DaySinceLastOrder : Positive skewed

Tenure : Positive skewed

3. Apakah sudah menemukan beberapa insight

- Dilihat dari feature Tenure, pengguna baru yaitu pengguna yang memiliki Tenure < 1 minggu lebih beresiko untuk churn.
- Nilai rata-rata satisfaction score tergolong biasa saja (3,06 out of 5) sehingga perlu adanya peningkatan layanan agar customer tetap berlangganan.
- Nilai rata- rata complain tergolong besar (0.28 out of 1) sehingga perlu adanya evaluasi layanan agar customer tetap berlangganan.
- Berdasarkan PreferredorderCat untuk kategori Laptop & Accessory memiliki complain yang lebih banyak dibandingkan kategori yang lain, sehingga beresiko untuk churn.
- Mayoritas customer churn berdasarkan gender dan marital status ialah laki-laki single sehingga perlu adanya spesial service untuk customer laki-laki yang masih single.

PEMBAGIAN TUGAS

1. **Lise Listianti**
   - Menulis notulensi mentoring Stage 1
   - Presentasi: Stage 0 dan Univariate Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Univariate Analysis dan Pembagian Tugas
2. **Ardilla Safitri**
   - Share screen dan mengedit perubahan saat diskusi setelah mentoring
   - Presentasi: Multivariate Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Multivariate Analysis dan Latar Belakang
3. **Arni Cici**
   - Presentasi: Descriptive Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Descriptive Analysis dan Business Insight
4. **Riska Diah**
   - Presentasi: Univariate Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Univariate Analysis dan EDA conclusion
5. **Jonse Kennedy**
   - Presentasi: Multivariate Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Multivariate Analysis dan menyatukan semua hasil
6. **Fakhry Husnul**
   - Presentasi: Descriptive Analysis
   - Laporan dan pembuatan slide: Menulis sebagian Descriptive Analysis

Rakamin Academy

TERIMA KASIH