Rakamin Academy

Laporan Final Project Byte Blazers

- Muhamad Faiz Widagdo
- Robiatul Adawiyah
- Chianti Ridhwan
- Lulu Safira
- Retno Debbi Yulisya
- Imam Luthfi
- Melliza Nastasia Izazi

Rakamin Academy

STAGE 0

Latar Belakang Masalah

TravelGuard+ adalah perusahaan yang menawarkan paket asuransi. Adanya pandemi COVID-19 menyebabkan penurunan minat pelanggan untuk membeli asuransi travel perjalanan ke luar negeri. Bahkan dari total 1987 pelanggan, hanya 710 orang (35.7%) yang memiliki asuransi travel.

Berdasarkan data yang diambil pada tahun 2019, TravelGuard+ ingin mengidentifikasi pelanggan yang sudah pasti akan membeli paket asuransi travel.

TravelGuard+ meminta tim Byte Blazers untuk membantu mengidentifikasi permasalahan tersebut agar penjualan terhadap apaket asuransi perusahaan dapat meningkat.

Goal

Meningkatkan penjualan paket asuransi perjalanan baru dalam satu tahun ke depan berdasarkan analisis pola pembelian pelanggan.

Objective

Membuat model machine learning yang dapat membantu bisnis travel asuransi ini memprediksi pelanggan mana yang akan membeli paket travel insurance.

Business Metrics

- Jumlah transaksi user pembelian paket asuransi travel

Role

Sebagai tim data scientist perusahaan PT Byte Blazers, kami diminta untuk menganalisis dan membuat machine learning yang dapat membantu TravelGuard+ meningkatkan penjualan paket asuransinya dengan memberikan rekomendasi kepada para pelanggan yang benar-benar akan membelinya.

Rakamin Academy

STAGE 1

EDA, Insights & Visualization

1.) Dataset Info

Menggunakan syntax df.info() untuk mengetahui informasi yang ada pada dataset.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1987 entries, 0 to 1986
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Unnamed: 0           1987 non-null   int64 
 1   Age                  1987 non-null   int64 
 2   Employment Type      1987 non-null   object
 3   GraduateOrNot        1987 non-null   object
 4   AnnualIncome         1987 non-null   int64 
 5   FamilyMembers        1987 non-null   int64 
 6   ChronicDiseases      1987 non-null   int64 
 7   FrequentFlyer        1987 non-null   object
 8   EverTravelledAbroad  1987 non-null   object
 9   TravelInsurance      1987 non-null   int64 
dtypes: int64(6), object(4)
memory usage: 155.4+ KB
```

- Dataset terdiri dari 1987 baris
- Terdapat 10 kolom
- Tidak ada missing values
- Semua tipe data sudah sesuai sehingga tidak ada yang diubah

EDA, Insights & Visualization

2.) Descriptive Statistics

Kami mengelompokkan kolom menjadi Categorical dan Numeric

```
[ ] # Pengelompokan kolom numeric dan kategori

    cat = ['Employment Type','GraduateOrNot','FrequentFlyer','EverTravelledAbroad']
    num = ['Age', 'AnnualIncome','FamilyMembers','ChronicDiseases','TravelInsurance']
```

```
[ ] df[num].describe()
```

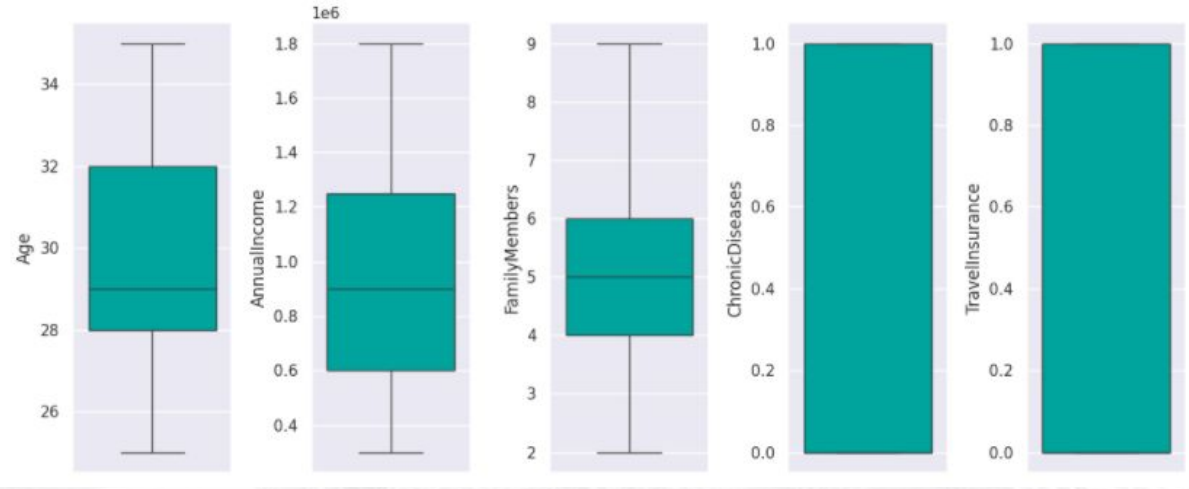
	Age	AnnualIncome	FamilyMembers	ChronicDiseases	TravelInsurance
count	1987.000000	1.987000e+03	1987.000000	1987.000000	1987.000000
mean	29.650226	9.327630e+05	4.752894	0.277806	0.357323
std	2.913308	3.768557e+05	1.609650	0.448030	0.479332
min	25.000000	3.000000e+05	2.000000	0.000000	0.000000
25%	28.000000	6.000000e+05	4.000000	0.000000	0.000000
50%	29.000000	9.000000e+05	5.000000	0.000000	0.000000
75%	32.000000	1.250000e+06	6.000000	1.000000	1.000000
max	35.000000	1.800000e+06	9.000000	1.000000	1.000000

Pengamatan Numerical:

Ada perbedaan antara nilai rata-rata dengan median dari ketiga kolom, yaitu `Age`, `AnnualIncome`, `FamilyMembers`, 'ChronicDiseases', dan 'TravelInsurance' namun tidak begitu signifikan

EDA, Insights & Visualization

2.) Univariate Analysis



Feature Numerical kami menggunakan boxplot dan displot. Berikut hasil pengamatan kami:

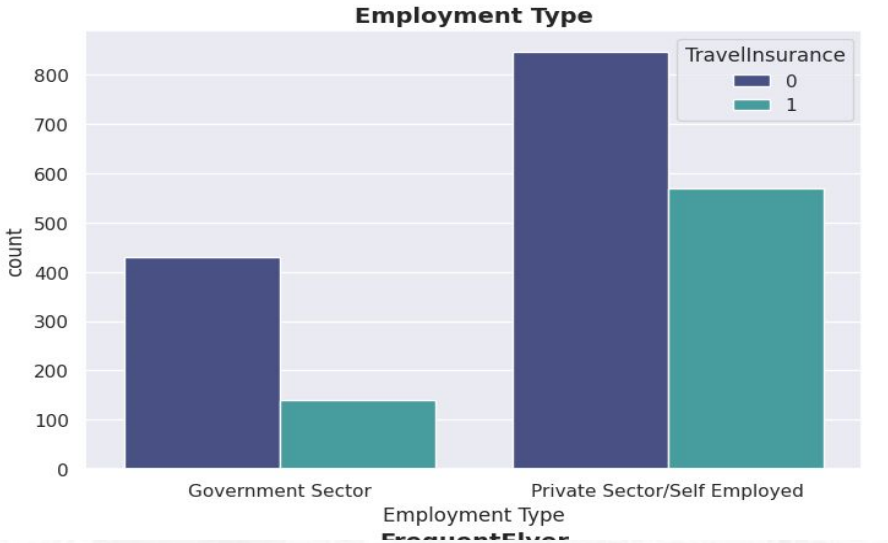
- Dapat dilihat dari boxplot disamping, kolom numerical tidak terdapat outliers
- Pada displot dapat dilihat kolom `numerical menunjukkan distribusi hampir normal, tidak ada skewness.

EDA, Insights & Visualization

Employment Type

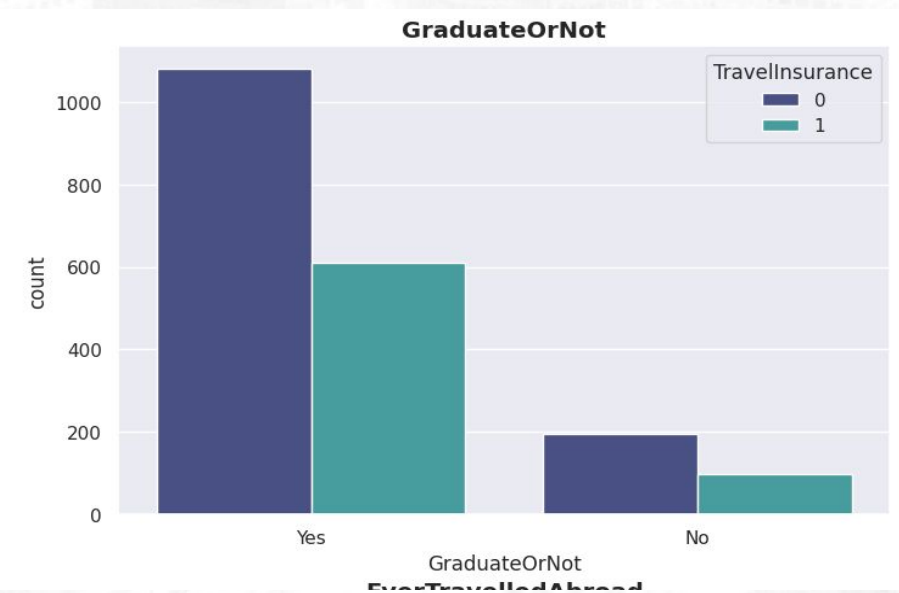
Baik customer yang bekerja di **Pemerintahan** maupun **Swasta** cenderung **tidak membeli Asuransi perjalanan**.

Namun rasio pembelian Asuransi di **Sektor Swasta lebih tinggi** (876 orang) dibandingkan **Pemerintah** (373 orang).



Graduate or Not

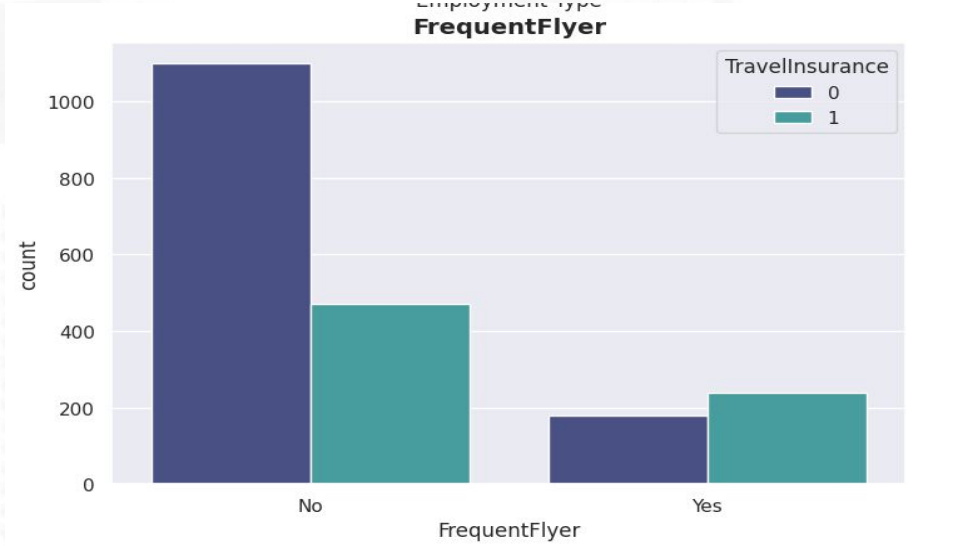
Tidak terdapat perbedaan yang signifikan dalam keputusan pembelian **Asuransi Perjalanan** antara customer yang tamat **sarjana atau** yang **tidak**.



EDA, Insights & Visualization

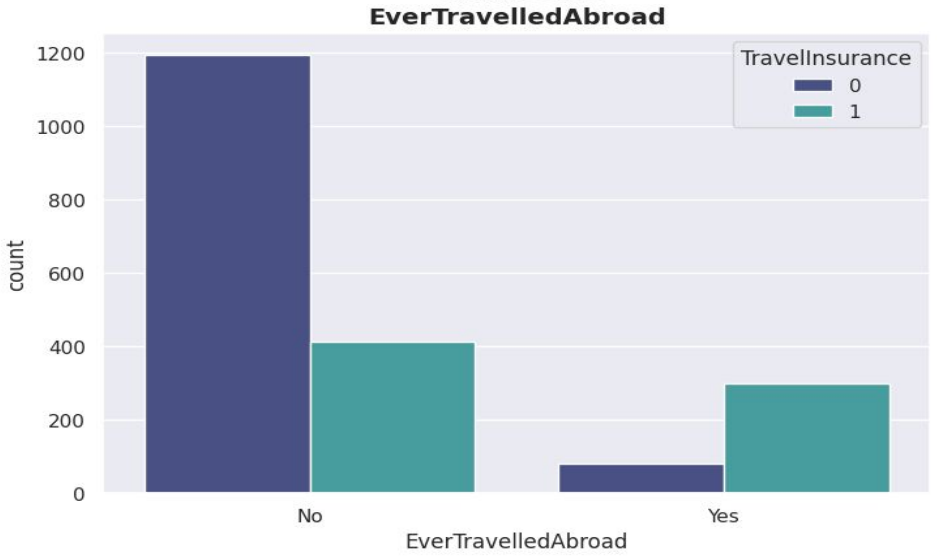
Frequent Flyer

Not Frequent Flyer, memiliki **potensi yang lebih tinggi** untuk membeli **Asuransi.**

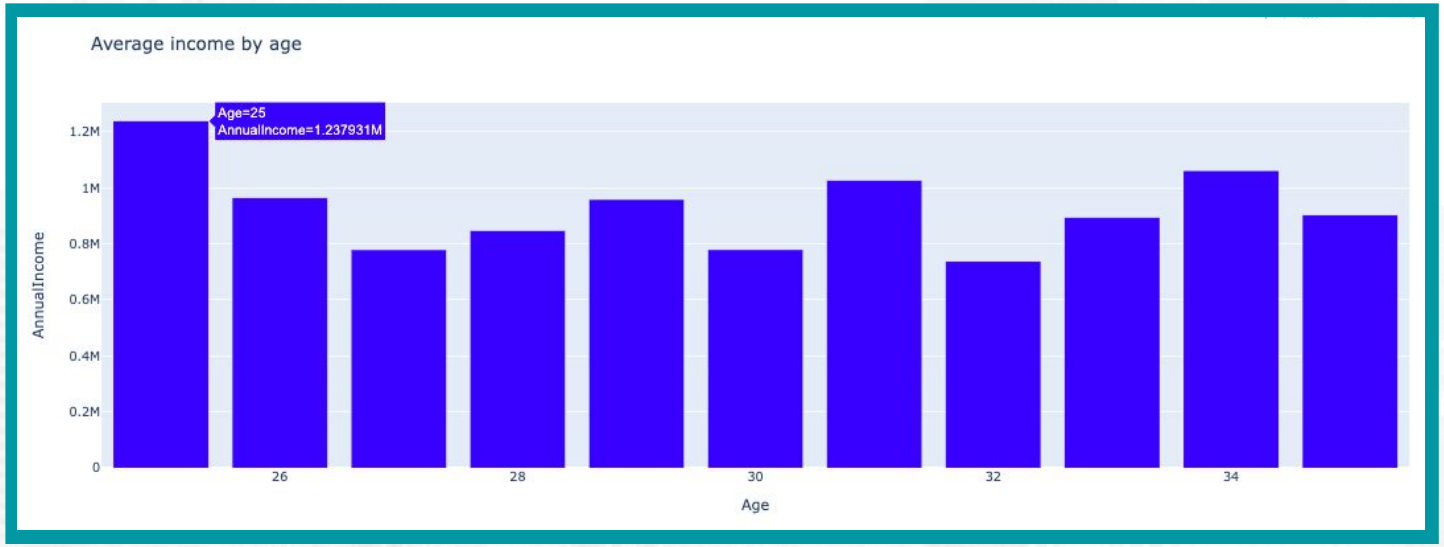


Ever Travelled Abroad

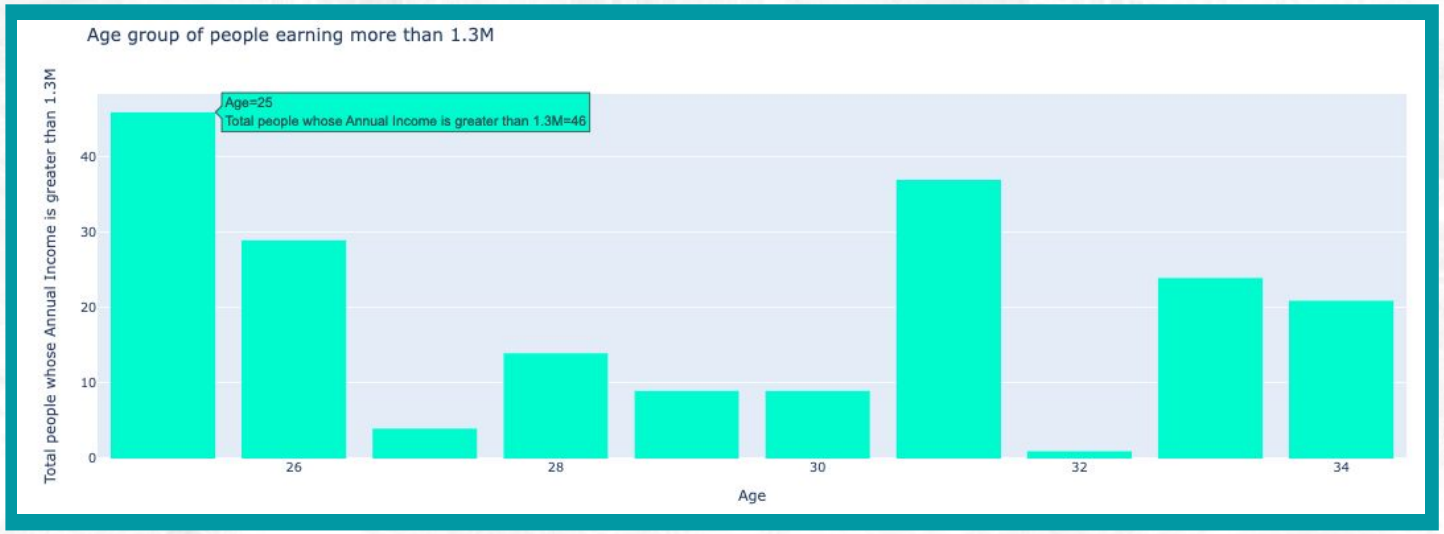
Pelanggan yang **pernah bepergian ke Luar Negeri cenderung membeli Asuransi.**



EDA, Insights & Visualization



Rata-rata **income tertinggi** berada di customer dengan umur **25 tahun**

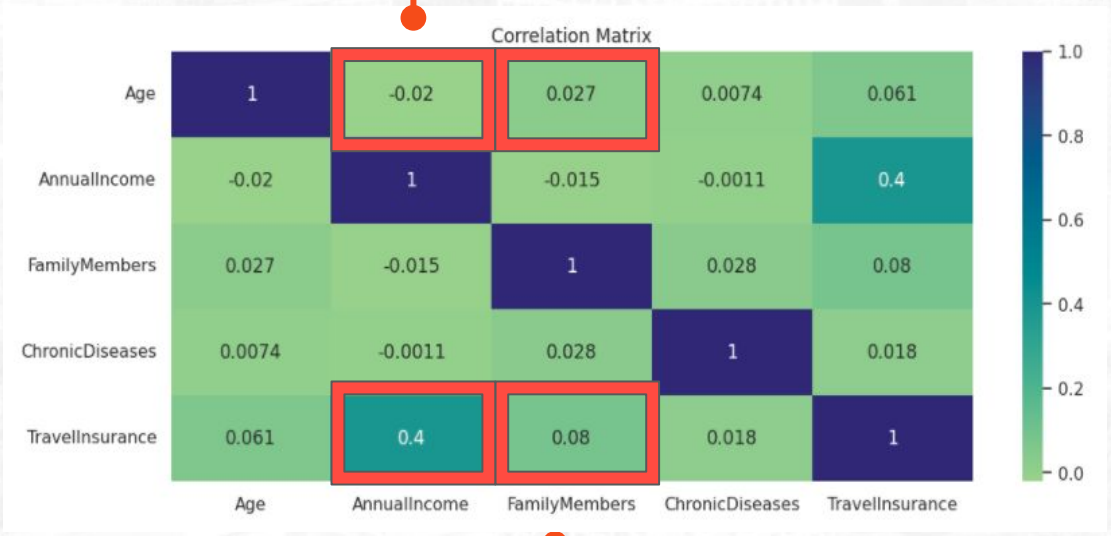


Dimana rata-rata tersebut menunjukkan income sebesar lebih dari 1.3M

EDA, Insights & Visualization

Analisis Multivariat merupakan metode statistik yang memungkinkan melakukan penelitian terhadap satu atau lebih dari dua variabel secara bersamaan. Untuk melihat korelasi feature kami menggunakan heatmap.

Age terhadap **Annual Income** dan Annual Income terhadap **Chronic Disease** memiliki **korelasi yang lemah dan negatif**



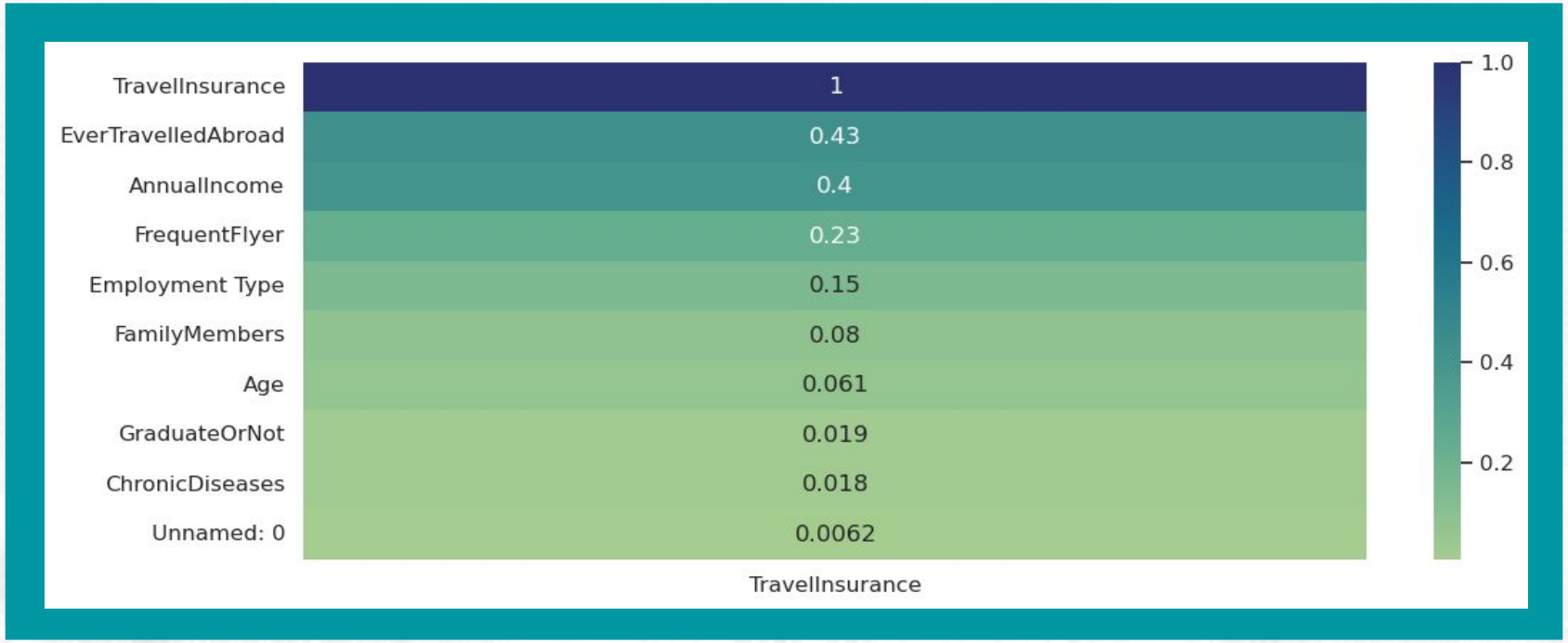
Age terhadap **FamilyMembers** dan FamilyMembers terhadap **Chronic Disease** memiliki **korelasi yang lemah dan positif**

Annual Income terhadap **TravelInsurance** memiliki **korelasi positif yang kuat**

FamilyMembers terhadap **TravelInsurance** memiliki **korelasi positif yang moderat**

EDA, Insights & Visualization

Korelasi masing-masing feature terhadap Travel Insurance



Feature EverTravelledAbroad, AnnualIncome, dan FrequentFlyer merupakan top 3 feature yang memiliki korelasi yang cukup tinggi terhadap Travel Insurance

EDA, Insights & Visualization

Business Insights

Mempertimbangkan untuk **mengarahkan** strategi pemasaran lebih khusus ke **pelanggan di sektor swasta**, mengingat rasio pembelian yang lebih tinggi di sektor ini.

Perusahaan dapat **menawarkan paket** asuransi keluarga khususnya untuk **keluarga beranggotakan 4 orang**

Perusahaan dapat **menawarkan paket member** travel insurance dengan **segmen annual income**

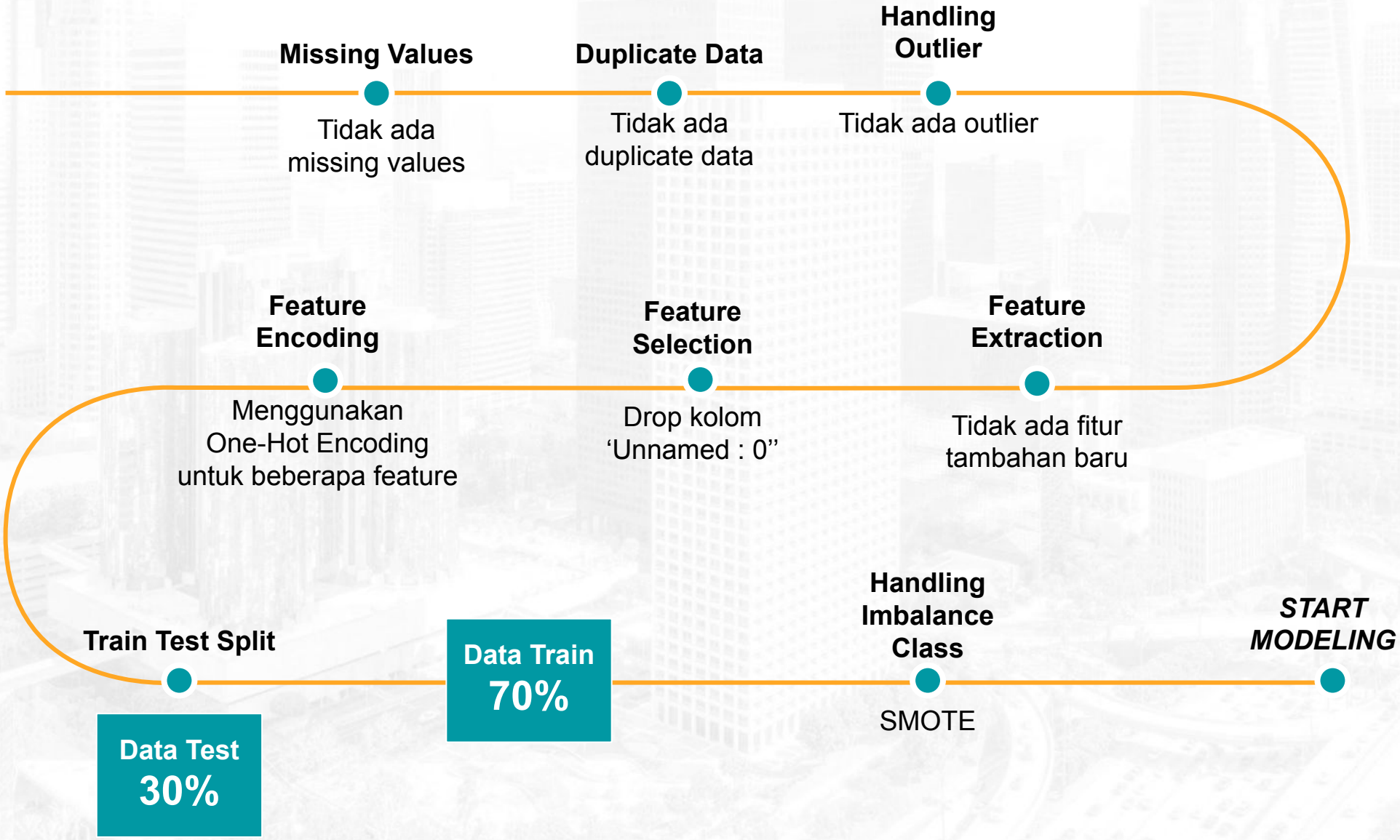
Mempertimbangkan strategi pemasaran yang **menawarkan berbagai promo** untuk paket wisata dengan menargetkan customer **yang tidak sering berpergian** dan **yang belum pernah bepergian ke luar negeri**.

Rakamin Academy

STAGE 2

Pre-Processing

Pre-Processing Flow



Handling Missing Value and Duplicate Data

Handling missing values dilakukan untuk mengatasi keberadaan nilai-nilai yang kosong atau hilang dalam suatu dataset. Hal ini sangat penting karena nilai yang hilang dapat memengaruhi hasil analisis dan model prediktif.

1.) Handling Missing Value and Duplicate Data

Untuk mendeteksi missing value digunakan .isnull().sum() kemudian menghapus baris yang mengandung nilai yang kosong dengan .dropna()

```
[ ] df.isnull().sum()

Unnamed: 0             0
Age                    0
Employment Type        0
GraduateOrNot          0
AnnualIncome           0
FamilyMembers          0
ChronicDiseases        0
FrequentFlyer          0
EverTravelledAbroad    0
TravelInsurance        0
dtype: int64
```

Setelah dilakukan Handling Missing Value **tidak terdapat data yang hilang** (*missing value*)

```
df.duplicated().any()

False
```

Tidak ada duplikat data.

Handling Redundant Data

Data Redundant adalah suatu kondisi ketika keberadaan informasi yang berlebihan atau berulang dalam suatu dataset.

Penanganan data redundan penting untuk menjaga kualitas data, mengoptimalkan penggunaan sumber daya, dan mencegah kesalahan analisis yang disebabkan oleh duplikasi atau ketidaksempurnaan dalam dataset. Penanganan data redundant adalah dengan melakukan pembersihan data dan menghapus atau menggabungkan informasi yang berulang.

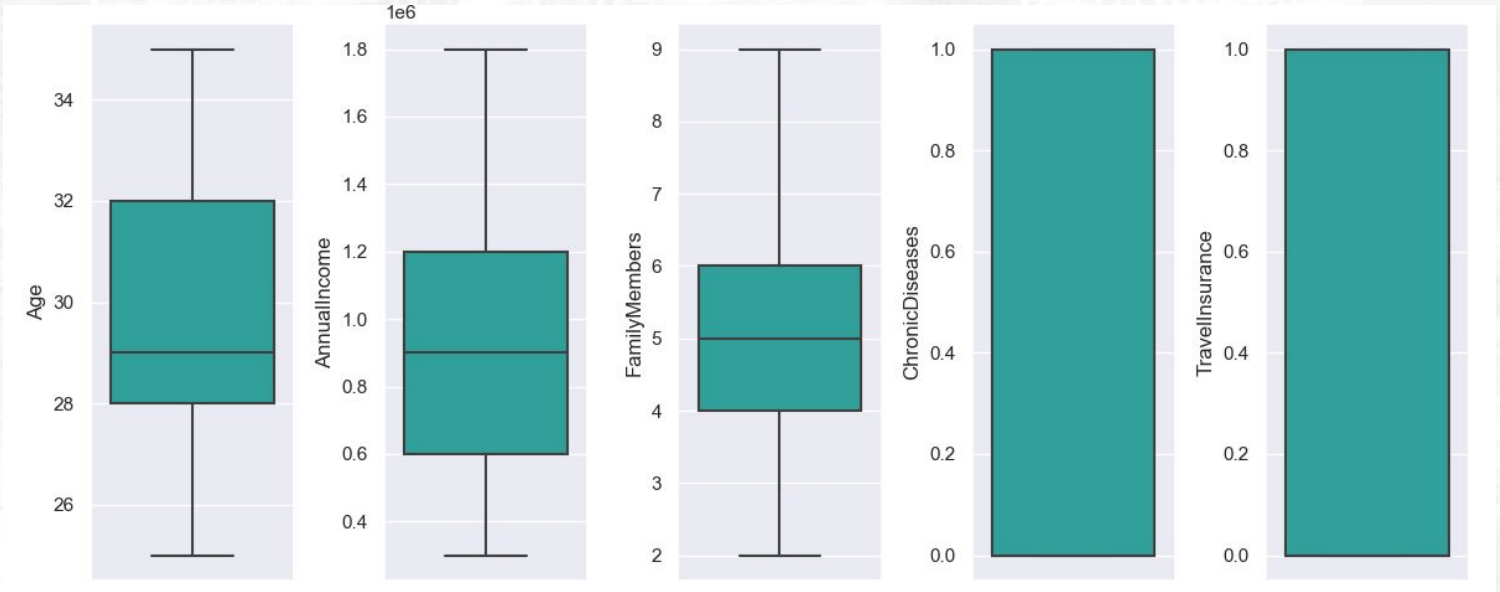
Pada dataset yang digunakan, kami tidak menemukan adanya data redundant yang perlu ditangani. Sehingga Handling Redundant Data tidak dilakukan.

Handling Outliers

Handling outliers dilakukan untuk mengatasi nilai-nilai yang sangat jauh atau tidak biasa dalam suatu dataset yang dapat mempengaruhi analisis statistik dan model prediktif, sehingga menangani mereka membantu mencegah kesalahan atau distorsi dalam interpretasi hasil analisis data.

```
# Boxplot untuk melihat outliers

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



Boxplot di samping menunjukkan bahwa **tidak terdapat outlier** pada data.

Feature Encoding

Kami melakukan feature encoding dengan one-hot encoding

```
[ ]  object_cols = df.select_dtypes(include=['object']).columns

     for col in object_cols:
        encoder = LabelEncoder()
        df[col] = encoder.fit_transform(df[col])
```

Split Data and Class Imbalance

Rakamin Academy

Split Data 70:30

```
[ ] # Train Test Split

    from sklearn.model_selection import train_test_split
    df.drop(columns=('Unnamed: 0'),inplace=True)

    X = df.drop(['TravelInsurance'], axis=1)
    y = df['TravelInsurance']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=69)
```

1. Pengecekan Class Imbalance

```
df['TravelInsurance'].value_counts()
```

```
0    766
1    483
Name: TravelInsurance, dtype: int64
```

Pada dataset ini kami meningkatkan jumlah sampel minoritas (TravelInsurance) dengan menciptakan sampel sintesis menggunakan Oversampling metode SMOTE

2. Handling imbalance data dengan resampling data

```
[ ] print("Original class distribution:")
    print(pd.Series(y_train).value_counts())

    # Create a SMOTE object with the desired sampling_strategy
    X_over_SMOTE, y_over_SMOTE = over_sampling.SMOTE().fit_resample(X_train, y_train)

    print("\nClass distribution after SMOTE:")
    print(pd.Series(y_over_SMOTE).value_counts())
```

```
Original class distribution:
0    904
1    486
Name: TravelInsurance, dtype: int64

Class distribution after SMOTE:
0    904
1    904
Name: TravelInsurance, dtype: int64
```

Rakamin Academy

STAGE 3

Machine Learning Modeling

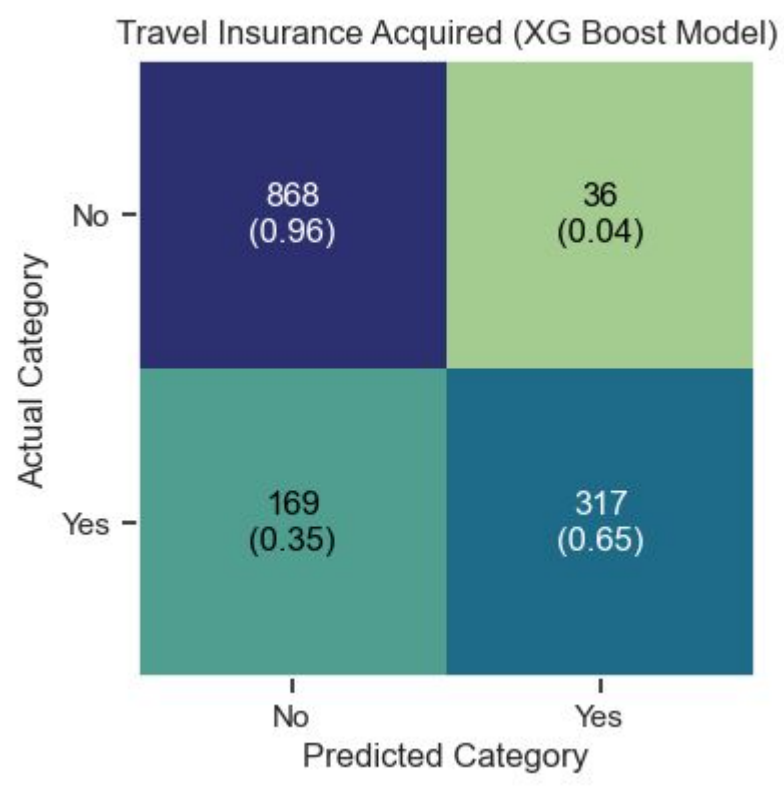
Model Evaluation

Kami melakukan pengujian pada empat model yaitu **Logistic Regression, Random Forest, Gradient Boosting, dan XG Boost** yang kemudian dilakukan hyperparameter tuning. **Metrics evaluasi yang kami lihat adalah Recall** karena kami ingin mengukur sejauh mana **model dapat mengidentifikasi semua kasus positif yang sebenarnya.**

Model	Before Hyperparameter Tuning		After Hyperparameter Tuning	
	Recall (Train)	Recall (Test)	Recall (Train)	Recall (Test)
Logistic Regression	0.55	0.58	0.47	0.52
Random Forest	0.86	0.67	0.59	0.60
Gradient Boosting	0.65	0.63	0.58	0.59
XG Boost	0.81	0.65	0.65	0.61

Berdasarkan analisis terhadap metrik recall dan metrik lainnya, **model XG Boost kemungkinan menjadi model terbaik** di antara keempat model yang telah di evaluasi.

Confusion Matrix



Kesimpulan:

- Model tampaknya **memiliki performa yang cukup baik pada data pelatihan dan data pengujian,** dengan akurasi sekitar 85.25% pada data pelatihan dan 80.90% pada data pengujian.

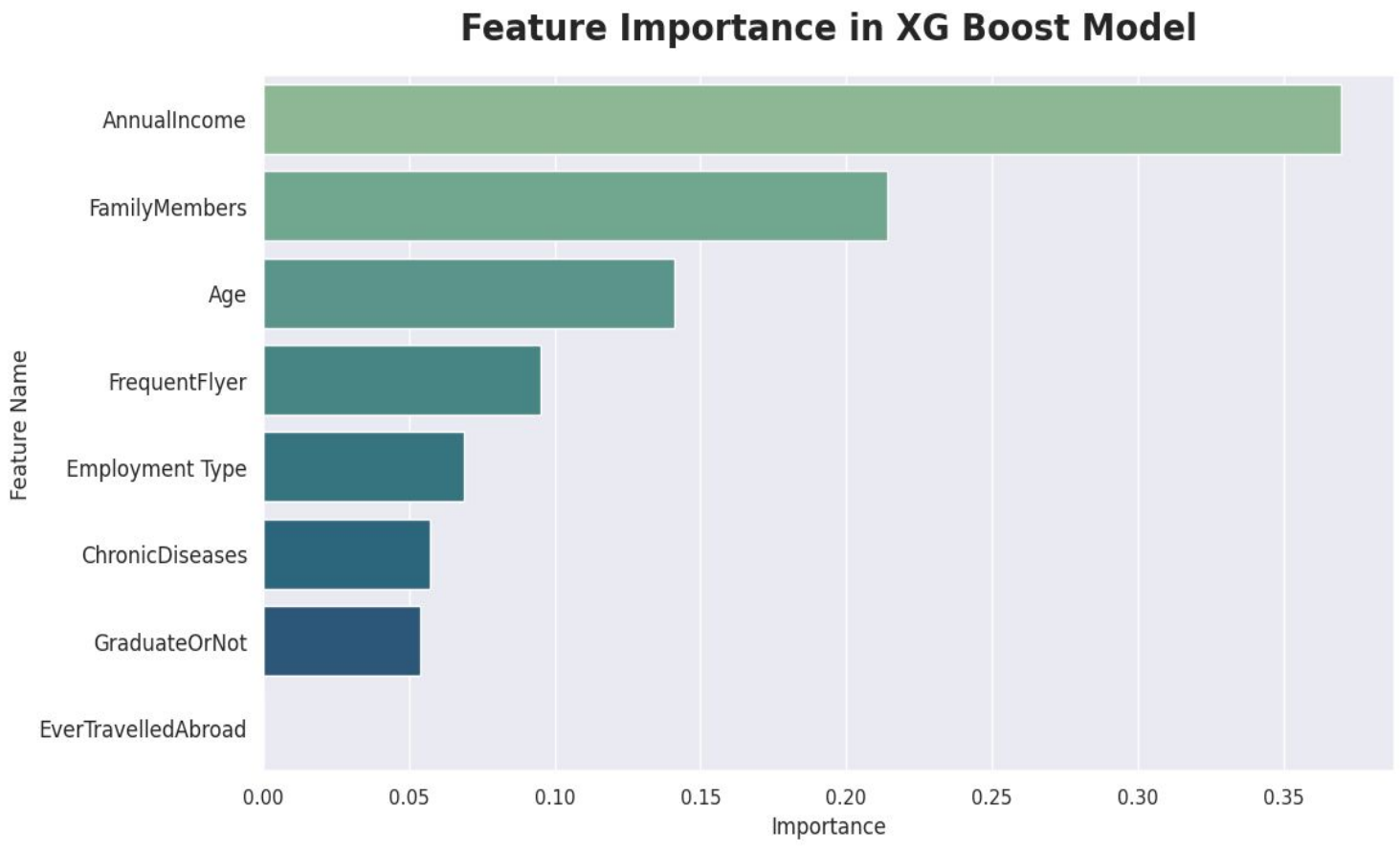
Recall = TP / TP + FN

= 317 / (317 + 169) * 100

= 65,22%

from 35.73% to 65.22% → Increase 29.49%

Feature Importance



Berdasarkan feature importance score disamping, dapat dilihat feature *AnnualIncome, FamilyMembers, dan Age* merupakan top feature importance yang dapat kita jadikan fokus untuk mendapatkan hasil prediksi yang akan membeli asuransi travel.

Selanjutnya feature ini dapat dilakukan landasan untuk feature selection pada iterasi selanjutnya.

STAGE 4

Business Recommendation

Business Recommendation

1. Pelanggan dengan kategori mid to high annual income berpotensi untuk membeli paket asuransi travel. Oleh karena itu tim marketing dapat membuat pendekatan strategis untuk menarik segmen pelanggan ini dengan memberikan perlakuan khusus yang bisa didapatkan oleh mereka (VIP Poin/Premium Poin)
2. Membuat pilihan paket untuk perjalanan domestic dan internasional
3. Mempertimbangkan pembuatan bundling promo untuk segmentasi pelanggan yang bepergian secara berkelompok (baik keluarga maupun instansi pemerintahan/swasta)
4. Melakukan Strategi Pemasaran untuk menarik pelanggan dalam kelompok Usia 25 tahun, karena mereka menunjukkan kemungkinan tertinggi untuk membeli paket Travel Insurance. Oleh sebab itu kelompok tersebut dilibatkan iklan dan promosi sebagai strategi pemasaran.
5. Mempertimbangkan pengelompokkan pelanggan berdasarkan pelanggan berdasarkan riwayat perjalanan, pendapatan tahunan dan anggota keluarga dalam paket Travel Insurance