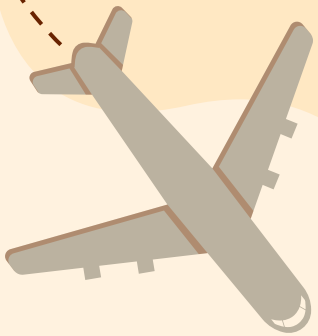
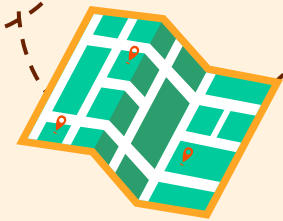




Travel Insurance Prediction

By Byte Blazers

Dataset!

Mengacu pada dataset TravelGuard Plus+, kami akan **memprediksi dan memberikan rekomendasi** untuk menentukan penggunaan paket asuransi perjalanan baru berdasarkan dari penjualan paket selama periode tahun 2019 dan dari calon pelanggan.

1 Latar Belakang Masalah

Adanya **penurunan** minat pelanggan untuk **membeli asuransi travel** perjalanan ke luar negeri **akibat adanya pandemi COVID-19.** Bahkan dari total 1987 pelanggan, **hanya 710 orang (35,69%) yang memiliki asuransi travel.**

2 Goal

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

3 Objective

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

4 Role

Sebagai tim data scientist perusahaan PT Byte Blazers, kami diminta untuk **memberikan rekomendasi Travel Insurance berdasarkan data** yang tersedia untuk **memberikan wawasan lebih dalam membimbing strategi pemasaran di masa depan**.

5 Business Metrics

Jumlah transaksi pembelian paket asuransi travel

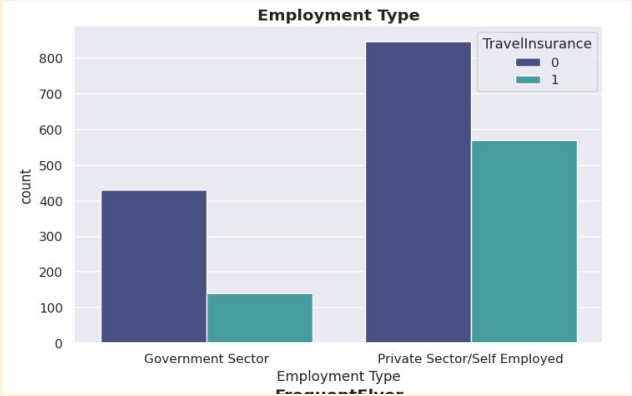
Feature Categorical,

kami coba untuk melihat pengaruh masing-masing feature terhadap pembelian Travel Insurance, berikut hasil interpretasinya:

Employment Type

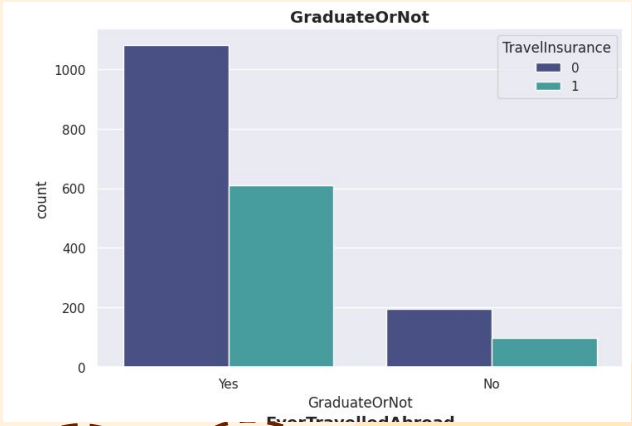
Baik customer yang bekerja di **Pemerintahan** maupun **Swasta** cenderung **tidak membeli Asuransi perjalanan**.

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**.

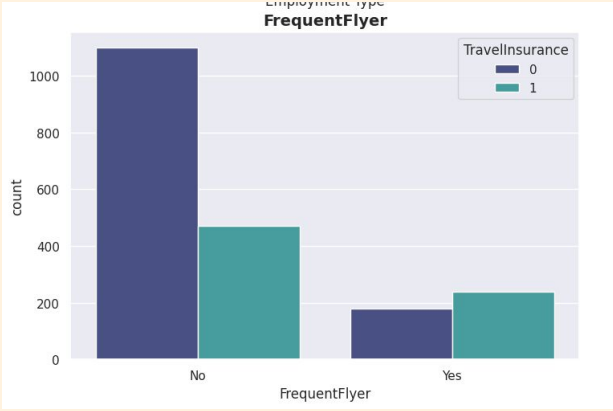


Feature Categorical,

kami coba untuk melihat pengaruh masing-masing feature terhadap pembelian Travel Insurance, berikut hasil interpretasinya:

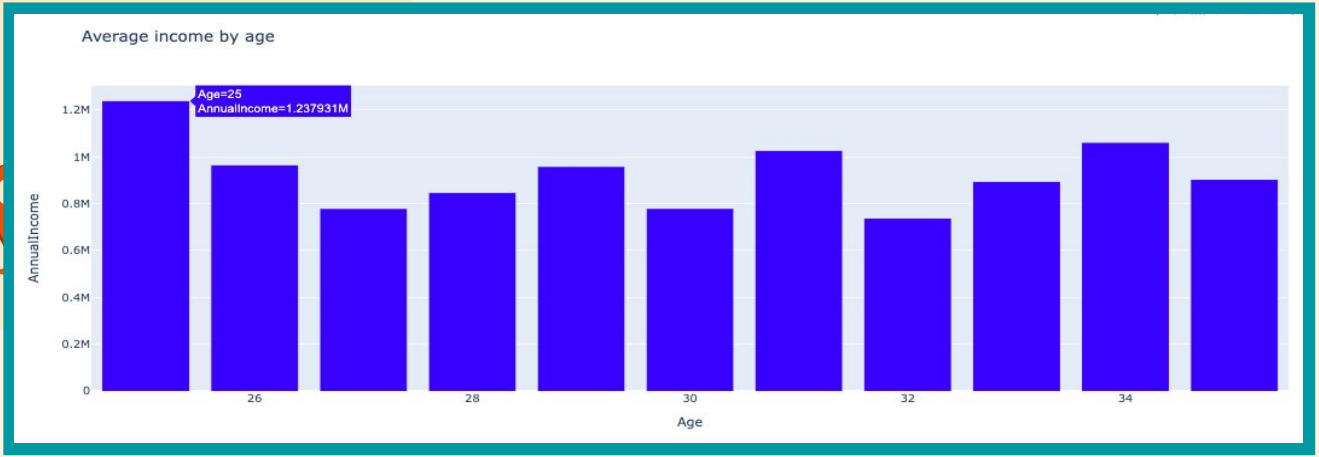
Frequent Flyer

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

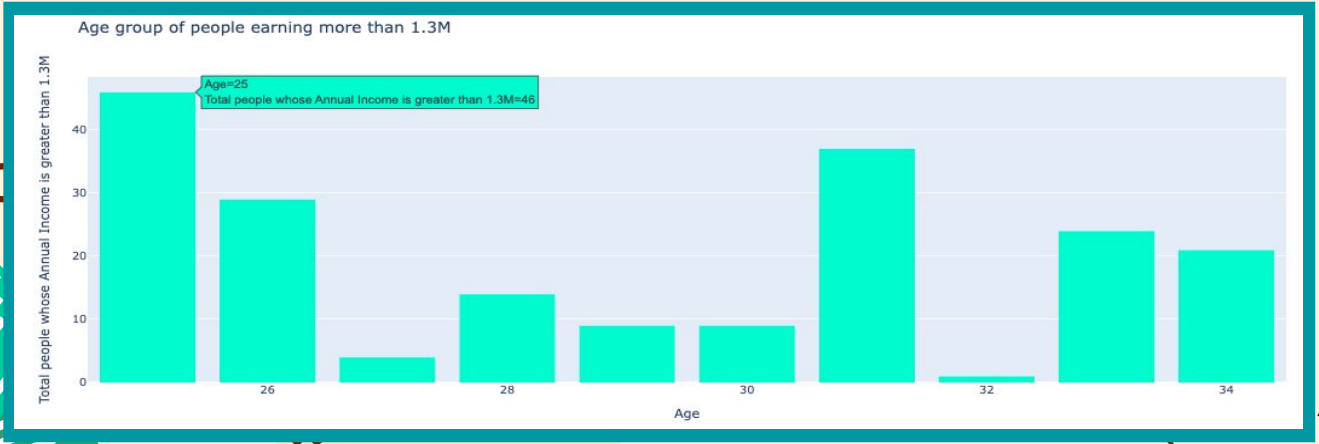


Ever Travelled Abroad

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



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

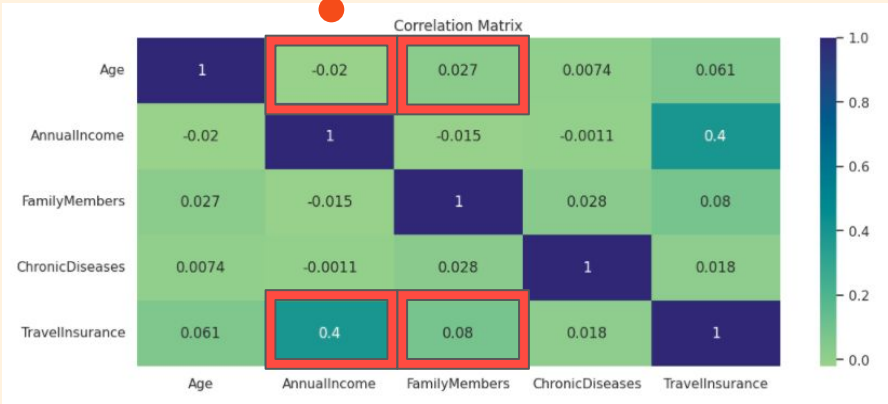


Dimana rata-rata tersebut menunjukkan income sebesar lebih dari 1.3M

Multivariate Analysis

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**

Multivariate Analysis

Kami pun melihat apakah faktor dari banyaknya Family Members mempengaruhi pembelian Asuransi Perjalanan, dan didaparkan data berikut:

1

Jumlah **Family Member** yang beranggotakan **4 orang mendominasi** sebagai **group yang membeli** Asuransi Perjalanan dibandingkan dengan jumlah anggota yang lain.

2

Semakin banyak anggota keluarga (>4 orang), peminat Asuransi Perjalanan **semakin menurun.**

Business Insight

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**.

Handling Missing Value

```
df.isna().sum()
```

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

- Dari hasil temuan kami, menunjukkan bahwa **tidak terdapat data yang hilang**.
- Sehingga pada data ini **tidak perlu dilakukan handling missing value**

Handling Duplicate Values

Mengatasi masalah ketika kita memiliki informasi yang sama atau sangat mirip dalam dataset.

- Dari hasil temuan kami, ditemukan bahwa **terdapat 738 data yang duplicate**.
- Maka dari itu kami **menghapus data duplicate** tersebut
- Sehingga perubahan datanya menjadi:

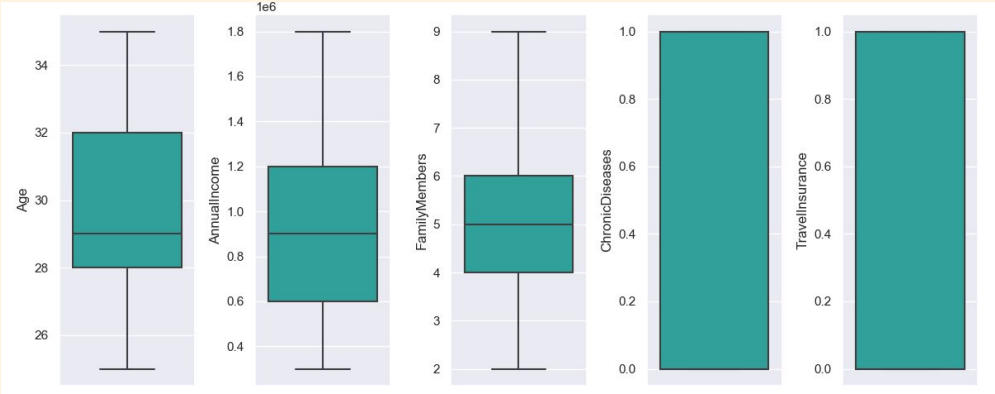
Data awal
1987 baris data

→

Data akhir
1249 baris data

Handling Outliers

Handling outliers dilakukan untuk mengatasi nilai-nilai yang sangat jauh berbeda dari nilai lainnya



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

Class Imbalanced

1. Pengecekan Class Imbalance

Tidak membeli Asuransi
766 orang
61%

:

Membeli Asuransi
483 orang
39%

Menunjukkan data yang digunakan tidak seimbang sehingga perlu dilakukan handling imbalanced data

2. Handling imbalance data dengan resampling data

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

```
Class distribution after SMOTE:
0    766
1    766
dtype: int64
```

Modelling Experiments

1.) Split Data Train and Test

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

2.) Train Model (Metrics: Recall)

Random Forest

Recall (Test Set): **0.75**
Recall (Train Set): 0.96

XG Boost

Recall (Test Set): **0.71**
Recall (Train Set): **0.92**

Logistic Regression

Recall (Test Set): **1.00**
Recall (Train Set): **1.00**

Gradient Boosting

Recall (Test Set): **0.51**
Recall (Train Set): **0.60**

Analisis dan kesimpulan hasil evaluasi metrics recall Model:

Random Forest:

- Model Random Forest juga menunjukkan kinerja yang baik. Terdapat sedikit perbedaan antara performa pada data uji dan data latih, namun perbedaan ini masih dapat diterima.

Logistic Regression:

- Performa yang sempurna pada data latih tetapi buruk pada data uji menunjukkan bahwa model tidak dapat melakukan generalisasi dengan baik ke data baru.

XG Boost:

- Model XG Boost memiliki kinerja yang sangat baik dengan recall sekitar 0.71 pada data uji dan 0.92 pada data latih. Hal ini menunjukkan bahwa model mampu mengidentifikasi sebagian besar data positif dengan baik tanpa mengorbankan performa pada data negatif.

Gradient Boosting:

- Terdapat perbedaan yang cukup signifikan antara performa pada data uji dan data latih, menunjukkan adanya overfitting.

Modelling Experiments

3.) Hyperparameter Tuning

Random Forest

Recall (Test Set): **0.62**
Recall (Train Set): **0.77**

XG Boost

Recall (Test Set): 0.66
Recall (Train Set): 0.82

Logistic Regression

Recall (Test Set): 1.00
Recall (Train Set): 1.00

Gradient Boosting

Recall (Test Set): 0.50
Recall (Train Set): 0.58

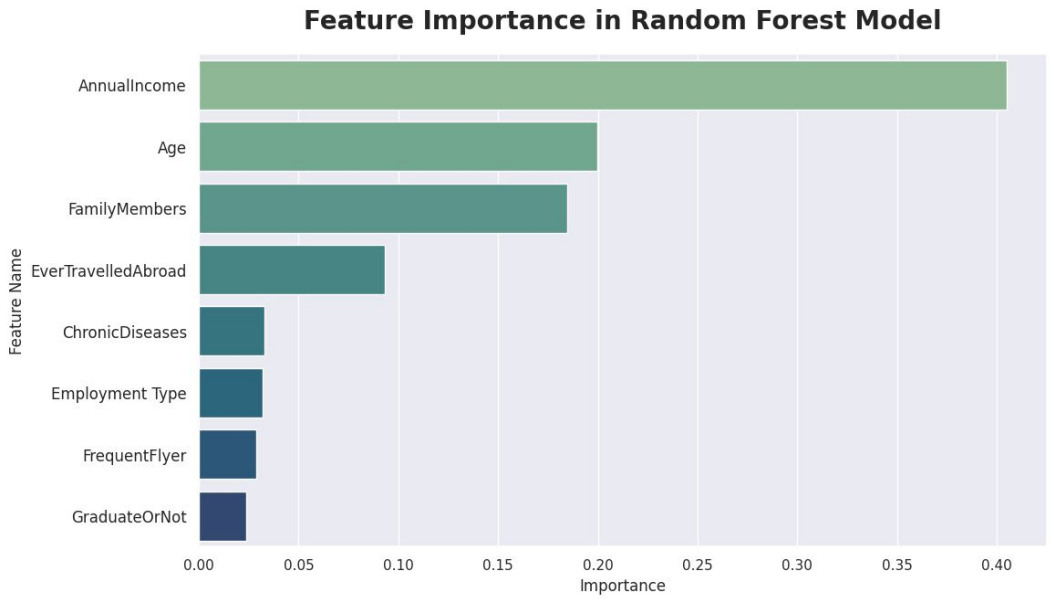
Setelah hyperparameter tuning, model Random Forest dan XG Boost menunjukkan kinerja yang lebih baik dalam hal recall dibandingkan dengan model Logistic Regression dan Gradient Boosting.

Namun, Random Forest menunjukkan performa yang lebih stabil dan sedikit overfitting dibandingkan dengan XG Boost.

Oleh karena itu, **Random Forest dapat dianggap sebagai model terbaik untuk digunakan dalam pengklasifikasi data ini.**

Modelling Experiments

4.) 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.**

Business Recommendation

PENGOPTIMALAN STRATEGI MARKETING

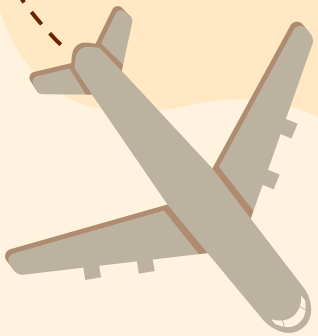
Analisis hasil dari model machine learning untuk mengidentifikasi pelanggan yang berpotensi membeli asuransi perjalanan baru dengan melihat tingkat recall digunakan untuk mengarahkan strategi pemasaran yang lebih terarah dan personalisasi ke tim marketing.

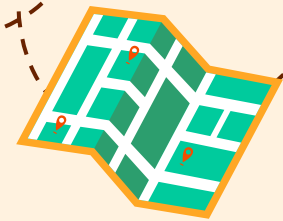
OPTIMASI PRODUK PAKET ASURANSI

Lakukan A/B testing untuk membandingkan performa produk dan paket asuransi yang baru dengan yang sudah ada. Ukur peningkatan penjualan dan respons pelanggan sebagai hasil dari perubahan ini.

PENINGKATAN KINERJA PENJUALAN

Pantau dan evaluasi kinerja tim penjualan secara berkala, dan ukur peningkatan penjualan yang dihasilkan dari implementasi strategi baru berdasarkan hasil model machine learning.





TERIMA KASIH

By Byte Blazers