Kelompok : 1
Stage : 0
Mentor : Fadilah Nur Imani
Pukul/ Tanggal : 20.00 WIB / January 18th 2023

Pembagian tugas di stage ini:

- Devriansyah Sya'ban : management tim, business insight & rekomendasi, management GIT
- Vira Diana : data cleansing, compile report & PPT
- Farih Afdhalul : feature engineering
- Dignu Akbar : feature engineering
- Kun Anggiar : data cleansing
- Laurenzius Julio : data cleansing & feature engineering
- Adinda Dita : data cleansing
- Ramadhani A : data cleansing

Poin pembahasan:

1. Handle bimodal data annual income
2. Feature transformation
3. Handle class imbalance
4. Feature selection.

Hasil Diskusi:

1. Handle bimodal data annual income
   Pada data annual income, setelah dilakukan distribution plot terindikasi bahwa datanya bimodal (memiliki beberapa modus). Akan tetapi setelah di cek lebih lanjut sebenarnya data tersebut hanya memiliki 1 modus saja. Terdapat beberapa data dengan frekuensi yang tinggi namun karena jumlahnya tidak sama persis, secara matematis tidak bisa dikatakan bimodal.

2. Feature transformation
   Feature transformation dengan log transformation akan dilakukan pada data yang positively skew.

Kelompok : 1
Stage : 0
Mentor : Fadilah Nur Imani
Pukul/ Tanggal : 20.00 WIB / January 18th 2023

Hasil Diskusi:

3. Handle class imbalance

   Untuk mengetahui adanya class imbalance, cukup dilihat dari labelnya saja.

   Presentase label dari dataset adalah 60% dan 30%. Sebenarnya perbedaannya tidak terlalu jauh, tapi juga tidak bisa dikatakan normal.

4. Feature selection

   Dari hasil analisa sementara, diperoleh beberapa feature yang akan digunakan untuk membuat model machine learning yaitu

   - Age
   - Annual income
   - Frequent flyer
   - Ever travelled abroad
   - Employment type

Kelompok : 1
Stage : 0
Mentor : Fadilah Nur Imani
Pukul/ Tanggal : 20.00 WIB / January 18th 2023

Tindak Lanjut:

1. Handle bimodal data annual income
   Data dapat terbukti bukan bimodal, jadi tidak perlu dilakukan drop data ataupun metode lainnya. Bisa ditunjukkan saja jika memang tidak bimodal
   Dicoba juga untuk bentuk data yang demikian, akan cocok dengan memakai algoritma yang mana saat membangun model.

2. Feature transformation
   Cek skewness setiap feature data, jika masih mendekati normal tidak perlu dilakukan feature transformation dengan log-transformation.

3. Handle class imbalance
   - Handle class imbalance dilakukan saat modelling,
   - Dicoba saja dulu apa adanya, jika performa class positifnya jauh dibandingkan performa class negatifnya, bisa dicoba pakai SMOTEN. Namun, jika performanya sudah bagus tidak perlu dilakukan metode untuk handle class imbalance.
   - Dicoba dulu dengan memakai algoritma sederhana seperti logistic regression atau Decission Tree untuk melihat performa model. Lalu dilakukan oversampling atau undersampling untuk dibandingkan hasilnya.
   - Bisa dicoba metode class weight, metode ini memberikan bobot ke modelnya. Biasanya performanya akan lebih bagus.

4. Feature selection
   - Pastikan sudah melakukan analisa dengan anggapan bahwa kita sudah masuk pada industri tersebut. Sehingga, secara knowledge sudah tau kira-kira harus memilih fitur apa saja
   - Jika feature yang dipilih merupakan data numerik, cek correlationnya.
   - Hati-hati dalam memilih feature dengan data kategorik. Karena data tersebut bukan real numbernya.
   - Agar lebih aman, gunakan hipotesis testing. Data kategorik pakai chi-square dan data numerical bisa pakai t-test.
   - Pemilihan fitur harus ada penjelasannya yang maksanse, plus didukung hasil pemodelannya