Homework Stage 2 - Preprocessing

Team:

1. Muhammad Gilang Mahardika
2. Muhammad Farhan Al Hafizh
3. Suny Guinesya Ardiansyah
4. Anuar Ali Syabana
5. Agus Setiana

Submission:

1. Report: https://docs.google.com/document/d/1QRvBeL-qKjZ-j4txNnEs1G96zhehRZ6yPK3g-KgvolA/edit?usp=sharing
2. Notebook: https://colab.research.google.com/drive/1KWNBCaXKu1FY1z-Gudw1H3zlOyz4zgII?usp=sharing

---

Features Cleansing

1. There are a total of 711 **duplicate** rows
    a. Sebuah baris akan ditandai sebagai duplikat ketika semua nilainya di seluruh kolom cocok dengan nilai dari baris lain atau baris lain dalam DataFrame.
    b. Baris duplikat biasanya dihapus setelah ditemukan karena mereka memiliki nilai yang sama dengan baris lain dan mungkin menyebabkan bias.
2. There are a total of 5 columns with **missing value**, This is **handled** by:
    a. Nilai yang hilang dalam kolom 'Administrative', 'Administrative_Duration', 'ProductRelated_Duration' diganti dengan 0. Asumsinya adalah pengunjung tidak mengunjungi halaman administrasi, sehingga tidak menghabiskan waktu di halaman tersebut.
    b. Asumsi pengunjung tidak menghabiskan waktu untuk menelusuri halaman terkait produk sehingga menghasilkan 0 detik di halaman tersebut.
    c. Kolom BounceRates diganti dengan nilai rata-ratanya, hal ini dilakukan karena rata-rata dan standar deviasinya tidak memiliki perbedaan nilai yang signifikan. Hal ini juga dapat berarti kurangnya nilai-nilai ekstrim.
    d. Kolom OperatingSystems diganti dengan nilai paling sering muncul.
3. **Drop unnecessary columns** resulted in the features Visitor “OperatingSystem” and “Browser” are judged to be irrelevant with the target feature.
4. **Handling Outliers**:
    a. **IQR**:
        i. Outliers are determined by the upper inner fence using the formula (Q3 + 1.5 * IQR).

ii. There are 10 features in the dataset that contain outliers. Among them, 2 features (Month and VisitorType) have values close to zero beyond their upper outlier threshold.
iii. The features with the highest percentages of outliers are TrafficType, BounceRates, and Administrative_Duration, with more than 10% of their values being outliers. This indicates that a significant number of data points fall outside the expected range of the majority of the data in these columns. As a result, the central tendency (mean, median) of these features is highly distorted.
iv. According to the National Institute of Standards and Technology, any values that exceed the upper inner fence of outliers (Q3 + 1.5*IQR) are considered to be mild outliers.

b. **Logarithmic Transformation**:
   i. Dari 10 fitur, hanya tersisa 5 fitur dengan keberadaan outlier, dengan 4 fitur (TrafficType, BounceRates, ExitRates, Region) masih memiliki persentase outlier yang cukup tinggi, dan 1 fitur (ProductRelated) mendekati nol persen.
   ii. Transformasi logaritmik berhasil mengurangi 5 dari 10 fitur menjadi sangat mendekati nol persen, yang kemudian dikecualikan.

c. **Winsorization**:
   i. Tidak ada fitur yang memiliki outlier setelah menggunakan winsorization.
   ii. Dengan menerapkan winsorization dengan batas persentil ke-5 dan ke-95, hal ini akan menggantikan nilai di bawah persentil ke-10 dan di atas persentil ke-90, serta di bawah persentil ke-15 dan di atas persentil ke-85, masing-masing.

5. **SMOTE**:
   a. Teknik oversampling yang membantu meningkatkan kinerja model machine learning dengan menyeimbangkan distribusi kelas, terutama berguna ketika kelas minoritas kurang terwakili.
   b. SMOTE bekerja dengan membuat contoh buatan dari kelas minoritas dengan interpolasi antara contoh kelas minoritas yang ada.

Feature Engineering

Feature Selection:

Total fitur dikeluarkan = 6
Total fitur digunakan = 12

1. Numerical features:
   A. Kriteria:
      - - Low cardinality (unique)
      - - No null values
      - - p-value < 0.05 (Chi2)

B. Hasil:
    - 1 fitur (TrafficType) dihapus karena nilai p-nya lebih dari kriteria ambang batas 0,05.
    - 2 fitur (Administrative, Administrative_Duration) dihapus karena nilai korelasinya kurang dari kriteria ambang batas 0,7.
    - 2 fitur (OperatingSystem, Browser) dihapus karena dianggap tidak relevan.
    - 1 fitur (revenue) dihapus untuk menjadi fitur target.

Feature Extraction:

1. Fitur "SpecialDay" dikategorikan menjadi 3 kategori (jauh, sedang, dekat), fitur ini berisi nilai yang menunjukkan seberapa dekat sebuah transaksi dengan hari istimewa terdekat.

Additional Features:

1. Usia pengguna.
2. Jenis kelamin pengguna.
3. Perjalanan penjelajahan: Halaman sebelumnya yang dikunjungi pengguna.
4. Tingkat klik: Rasio elemen yang diklik oleh pengguna.