Kelompok 3B : Insight Miner
Nama Kelompok :
1. Purwantiyo
2. Anisa Agustiana
3. Gideon Kurniawan Sugiarto
4. Hendra Hermawan
5. Faradila Suwandi
6. Widya Angela
7. Regina Aprilia Sembiring
8. Yolan Faiz Jamahsyari

1. Data Cleansing

1a. Handle Missing Value

untuk menangani data missing disini hal pertama yang dilakukan adalah mengecek data missing kemudian menghitung persentase data missing tersebut ternyata rata rata data missing berada di bawah 10% yang artinya tidak signifikan lalu untuk menghadel data missing tersebut digunakan metode fillna dengan model mean dikarenakan data yang missing merupakan data numerik.

1b. Handle Duplicate Data

disini tidak ada hal yang perlu ditangani karena setelah dicek ternyata data yang digunakan tidak memiliki data duplikat yang berarti data tersebut dapat langsung digunakan

1c. Handle Outlier

Menangani outlier dalam data adalah langkah penting dalam analisis data untuk memastikan bahwa nilai-nilai ekstrim tidak terlalu mempengaruhi hasil. langkah pertama untuk menangani outlier yaitu perlu mengidentifikasi outlier dalam kumpulan data menggunakan metode statistik seperti “zscore” atau metode IQR (interquartile Range). Namun, Terkadang, Outlier sebenarnya mengandung informasi penting . Outlier dapat mengidentifikasi situasi atau kondisi yang unik atau tidak biasa dalam data, dan menghapusnya dapat menghilangkan wawasan yang berharga.

1d. Feature Transformation

Future Transformation adalah langkah penting dalam pra-pemrosesan data dan dapat melibatkan beberapa teknik disini melakukan log transformation pada kolom yang memiliki distribusi data yang condong.

1e. Feature encoding

Dalam tahapan Feature encoding kita mengonversi variabel kategorikal (atau non-numerik) dalam dataset menjadi bentuk numerik sehingga dapat digunakan oleh algoritma machine learning. Beberapa algoritma machine learning hanya dapat bekerja dengan data numerik, oleh karena itu variabel kategorikal perlu diubah ke dalam bentuk numerik.

1f. Handle class imbalance

Dalam proses handle class imbalance kita melakukan penyeimbangan untuk data data yang tidak seimbang jumlah atau nilainya. Ini biasanya dilakukan pada data yang memiliki perbedaan yang tidak seimbang antara kelas. Tujuannya adalah untuk mengatasi masalah ketidakseimbangan ini agar model machine learning dapat mempelajari pola dari kelas

minoritas dengan baik. Untuk kasus ini kami menggunakan metode smote untuk Membuat sampel sintetis dari kelas minoritas untuk meningkatkan jumlahnya.

2. Feature Engineering

a. Feature Selection (membuang feature yang kurang relevan atau redundan)
dimana pada feature ini kita memilih fitur-fitur untuk modeling dan mengidentifikasi fitur yang terbaik.

b. Feature extraction (membuat feature baru dari feature yang sudah ada)
Hasil dari feature ekstraksi dari feature yang sudah ada, didapatkan hasil 4 fitur baru yakni:

   a. TotalSpending
   hasil penggabungan fitur "OrderAmountHikeFromlastYear" dan "OrderCount" untuk membuat fitur yang mengukur total pengeluaran pelanggan dalam setahun. Feature ini dapat memberikan wawasan tentang tingkat pengeluaran pelanggan.

   b. AverageOrderValue
   Hasil perhitungan rata-rata nilai pesanan dengan membagi "OrderAmountHikeFromlastYear" dengan "OrderCount". Feature ini ni dapat memberikan informasi tentang seberapa besar setiap pesanan yang ditempatkan oleh pelanggan.

   c. CashbackPercentage
   Hasil perhitungan persentase uang kembalian berdasarkan "CashbackAmount" dibandingkan dengan total pengeluaran ("OrderAmountHikeFromlastYear").

   d. FrequencyOfOrders
   Fitur yang mengukur seberapa sering pelanggan melakukan pesanan dengan membagi "OrderCount" dengan "Tenure." Feature ini berguna untuk memberitahu tentang kebiasaan pemesanan pelanggan.

Namun setelah dilihat korelasi antara feature baru dan feature lama menggunakan heatmap, terdapat beberapa korelasi sangat kuat, yakni >0,7. Hal ini ditakutkan akan menimbulkan redundan, sehingga harus diambil tindakan untuk menghilangkan salah satu feature yang memiliki kolerasi sangat kuat tersebut. Setelah melihat korelasi antara feature baru dengan target/'Churn', dapat diambil kesimpulan bahwa hanya 2 feature baru yang dapat ditambahkan, yakni feature 'AvarageOrderValue' dan 'FrequencyOfOrders'

c. Fitur Tambahan
Pada bagian ini, kami memberikan ide fitur-fitur yang belum ada di list, namun dianalisis dapat memberikan modelling yang lebih baik apabila fitur tersebut diadakan. Fitur-fiturnya yaitu Penghasilan rumah tangga, Frekuensi pembelian per periode waktu tertentu, Jumlah uang yang dihabiskan per periode waktu tertentu, Harga produk yang relatif dibeli pelanggan, Jarak ke toko terdekat, umur, dan Rata-rata waktu pengiriman.
Mengapa fitur-fitur tersebut perlu ditambahkan? Alasannya:

- Harga produk yang relatif dibeli pelanggan: jika harga produk naik lebih mahal dibandingkan dengan pesaing, customer mungkin akan mencari alternatif yang lebih terjangkau. Harga yang lebih rendah atau penawaran yang lebih baik dari pesaing dapat mempengaruhi customer untuk churn.

- Jarak ke Toko Terdekat: Jika data jarak customer ke toko fisik dekat, kemungkinan churn pelanggan akan semakin besar karena pelanggan mungkin lebih sering berkunjung ke toko terdekat daripada membeli di e-commerce sehingga pelanggan kemungkinan akan churn.
- Umur: Jika informasi umur pelanggan tersedia, mungkin fitur ini dapat menentukan apakah ada hubungan antara usia pelanggan dan perilaku churn.
- Rata-rata waktu pengiriman: Semakin lama, customer mungkin merasa tidak puas sehingga dapat menjadi faktor churn. Sebaliknya, jika waktu pengiriman sesuai dengan harapan, maka customer merasa puas dan dapat membantu mengurangi tingkat churn dengan menjaga kesetiaan customer.