

IS184943 - PENGGALIAN DATA

TUGAS KELOMPOK I

EKSPLORASI DAN PRAPROSES DATA

Tugas kelompok pertama berkaitan dengan eskplorasi data (*Exploratory Data Analysis/EDA*) dan praproses data terhadap sebuah data mengenai prediksi kemungkinan seorang kastemer akan melakukan tindakan loncatan (*customer churn prediction*) menggunakan python. Gunakan referensi dari berbagai sumber, termasuk catatan yang diberikan dalam kuliah dan tutorial. Hasil eskplorasi dan praproses data yang ada lakukan akan digunakan dalam tugas kelompok berikutnya (Klasifikasi).

A. Deskripsi data

Rangkuman

Data yang terdiri dari 28.382 baris dan 22 kolom merupakan data sebuah Bank yang bermaksud untuk mempertahankan para kastemernya untuk produk perbankan, yaitu saldo tabungan (*saving accounts*). Pihak bank menginginkan untuk mengidentifikasi apakah seorang kastemer kira-kira akan melakukan *churn* terhadap saldo tabungannya. Dalam hal ini, *churn* dapat dipandang sebagai tindakan yang tidak diharapkan sehingga saldo tabungannya tersisa di bawah saldo minimum.

Dalam set data, anda diberikan informasi mengenai kastemer, seperti usia, jenis kelamin, data demografi beserta beberapa data transaksi perbankan yang dilakukan oleh kastemer. Setiap baris data merepresentasikan seorang kastemer, di mana setiap kolom data berisikan atribut yang terkait dengan demografi kastemer dan juga berbagai transaksi yang pernah dilakukan oleh kastemer. Untuk ini, terdapat beberapa variabel/atribut dalam set data yang dapat dikelompokkan menjadi 3 kategori seperti berikut.

Demographic information about customers:

- customer_id** - Customer id
- vintage** - Vintage of the customer with the bank in number of days
- age** - Age of customer
- gender** - Gender of customer
- dependents** - Number of dependents
- occupation** - Occupation of the customer
- city** - City of customer (anonymised)

Bank Related Information for customers:

- customer_nw_category** - Net worth of customer (3:Low 2:Medium 1:High)
- branch_code** - Branch Code for customer account
- days_since_last_transaction** - No of Days Since Last Credit in Last 1 year

Transactional Information:

- current_balance** - Balance as of today
- previous_month_end_balance** - End of Month Balance of previous month
- average_monthly_balance_prevQ** - Average monthly balances (AMB) in Previous Quarter
- average_monthly_balance_prevQ2** - Average monthly balances (AMB) in previous to previous quarter
- percent_change_credits** - Percent Change in Credits between last 2 quarters
- current_month_credit** - Total Credit Amount current month
- previous_month_credit** - Total Credit Amount previous month
- current_month_debit** - Total Debit Amount current month
- previous_month_debit** - Total Debit Amount previous month
- current_month_balance** - Average Balance of current month
- previous_month_balance** - Average Balance of previous month
- churn** - Average balance of customer falls below minimum balance in the next quarter (1/0)

B. Tugas

1. Lakukan eksplorasi data dari berbagai perspektif untuk memahami karakteristik data agar anda mempunyai persepsi yang baik terhadap data. Gambakan hasil eksplorasi dalam berbagai bentuk grafik/*chart* yang menurut anda paling sesuai untuk menggambarkan karakteristik data secara komprehensif dan mudah dipahami (anda dapat menggunakan library visualisasi data yang populer dalam python, yaitu **Seaborn** yang pernah dijelaskan dalam tutorial).

Analisis eksploratori yang harus dilakukan paling tidak meliputi analisis **univariate** (analisis dari setiap atribut) dan analisis **bivariate** (analisis korelasi) antara setiap atribut independen dan atribut dependen (atribut kelas, yaitu *churn*). Khusus untuk analisis bivariat, paling tidak ada harus melakukan analisis terkait dengan beberapa hipotesis berikut:

- a. Apakah kastemer lama (*vintage customer*) cenderung untuk melakukan churn?
 - b. Apakah kastemer dengan rata-rata saldo tertinggi akan cenderung untuk melakukan churn?
 - c. Apakah kastemer yang mengalami penurunan saldo bulanan akan cenderung untuk melakukan churn?
 - d. Apakah kastemer perempuan mempunyai kecenderungan yang rendah untuk melakukan churn?
 - e. Apakah kastemer muda akan mempunyai kecenderungan melakukan churn?
 - f. Apakah kastemer yang dengan penghasilan kecil akan cenderung melakukan churn?
 - g. Apakah kastemer yang memiliki tanggungan keluarga cenderung untuk melakukan churn?
 - h. Apakah kastemer dengan rata-rata jumlah tanggungan keluarga kurang dari 4 cenderung untuk melakukan churn?
 - i. Apakah kastemer yang melakukan transaksi terakhir lebih dari 6 bulan lalu mempunyai kecenderungan untuk mempunyai kemungkinan churn yang lebih tinggi?
 - j. Apakah ada kemungkinan kota dan cabang dari bank dengan jumlah nasabah yang kecil akan cenderung melakukan churn
2. Lakukan praproses data yang menurut anda diperlukan sebelum dilakukan proses penggalian data (dalam tugas berikutnya). Praproses antara lain meliputi pengecekan setiap tipe data atribut dan juga penyesuaian tipe data (jika tipe data yang ditentukan oleh python secara otomatis dianggap kurang tepat atau tidak dikenali oleh python yaitu tipe data *object*). Selain itu juga perlu dilakukan praproses terkait dengan *missing value* dan data pencilan (*oulier*).

C. Laporan dan Batas Waktu

- Laporan ditulis pada kertas berukuran A4 dengan spasi tunggal. Laporan dalam format PDF diserahkan per kelompok dan diunggah dalam menu "Assignment" pada aplikasi Teams **paling lambat pada tanggal 14 November 2020 pukul 16:00 WIB** (hanya satu orang dari setiap kelompok yang harus menyerahkan laporan).
- Penilaian akan didasarkan pada aspek: sistematika penulisan dan kelengkapan laporan (15%), kejelasan uraian laporan termasuk tata tulis (15%), kejelasan metode dan uraian hasil eksplorasi data (40%), dan kejelasan metode dan uraian praproses data (30%).
- Selain laporan tugas, setiap kelompok juga harus mengunggah *script program python* (saved file dalam *Jupyter Notebook*) dalam format <nama-file>.ipynb.
- Isi laporan dan script program yang mengindikasikan adanya plagiarisme tidak akan dinilai.

-----oooOooo-----