**[Lanjut besok]**

**LAPORAN FINAL PROJECT**

Data Science

(Model Maestro)

Anggota:

* Mario Trianto
* Rendy Budi Kusuma
* Muhammad Niko Dwi Putranto
* Hilda Amalia
* Prambayu Rehyta Muryandari
* Ni Komang Tri Anjani
* Abu Bakar Sidiq

1. **Pendahuluan**
   1. Latar Belakang

Bank memiliki beragam sumber pendapatan, salah satunya adalah deposito berjangka (term deposit). Deposito berjangka merupakan investasi tunai yang disimpan di lembaga keuangan dengan tingkat bunga tetap selama jangka waktu tertentu. Produk ini sangat penting bagi bank karena memberikan aliran dana yang stabil dan mendukung operasional serta pertumbuhan bisnis.

Bank menggunakan berbagai strategi pemasaran untuk menjual produk deposito ini, seperti pemasaran melalui email, iklan, pemasaran telepon dan pemasaran digtal. Pemasaran telepon (telemarketing) masih menjadi salah satu cara yang paling efektif untuk mencapai calon pelanggan, karena interaksi langsung melalui telepon memungkinkan bank untuk memberikan penjelasan rinci dan menjawab pertanyaan pelanggan secara real-time. Namun, pemasaran telepon juga memerlukan investasi besar, terutama karena melibatkan menyewaan pusat panggilan (call center) dalam jumlah besar. Karena itu, penting bagi bank untuk mengidentifikasi pelanggan yang paling mungkin untuk berlanggananan deposito sebelum melakukan campaign telepon, sehingga sumber daya dapat dioptimalkan.

Dalam konteks ini, proyek ini menggunakan data campaign pemasaran telepon langsung dari sebuah lembaga perbankan di portugal untuk memprediksi apakah seorang pelanggan akan meyetujui untuk membuka deposito berjangka. Dengan menggunakan teknik machine learning, kami bertujuan untuk mengembangkan model prediksi yang dapat membantu bank menargetkan pelanggan dengan lebih efektif, mengurangi biaya campaign pemasaran, dan meningkatkan tingkat konversi pelanggan.

Berdasarkan data yang digunakan dalam proyek ini, diketahui bahwa dari keseluruhan data yang ada, hanya 11,7% pelanggan yang melakukan pembukaan deposit.

* 1. Goal and Objective

Berdasarkan masalah tersebut, diketahui bahwa proyek ini bertujuan untuk meningkatkan jumlah direct customer yang membuka deposito untuk meningkatkan revenue perusahaan. Untuk lebih jelasnya, tujuan utama dari proyek ini adalah untuk membuat sebuah model rekomendasi untuk membantu menentukan nasabag yang berpeluang tinggi akan melakukan deposit dan meningkatkan dorect customer. Sehingga, bussiness metrics dalam proyek ini adalah :

* + 1. Convertion Rate : Yaitu jumlah justomer yang membuka taungan deposit
    2. Cntact rate (?)
    3. ROI (?)

1. **Exploratory Data Analysis (EDA)**

EDA dilakukan untuk memahami distribusi data, memeriksa kualitas data, dan menemukan pola-pola penting yang mungkin relevan untuk memprediksi apakah pelanggan akan subscribe atau tidak.

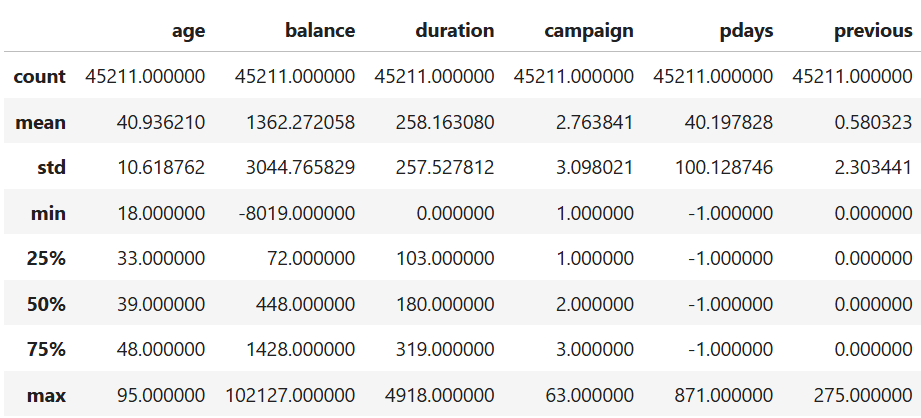
1. About Dataset

Data ini terdiri dari 45.211 entri baris dan 17 kolom yang berisi informasi demografi dan historis campaign sebelumnya. keterangan feature berupa:

1. Berhubungan dengan informasi customer :
   * + 1. age (numeric) : Berisi data umur pelanggan
       2. job (categorical) : tipe job pelanggan
       3. marital (categorical) : status pernikahan pelanggan
       4. education (categorical) : status pendidikan pelanggan
       5. default (Categorical) : (y/n) status kredit pelanggan
       6. balance (numeric) : rata-rata penghasilan pelanggan dalam setahun
       7. housing (Categorical) : (y/n) status cicilan rumah pelanggan
       8. loan (Categorical) : (y/n) status cicilan pelanggan
2. Berhubungan dengan historis campaign sebelumnya :
3. contact (categorical) : tipe komunikasi kontak terakhir
4. day (numerical) : tanggal kontak terakhir
5. month (Categorical) : bulan kontak terakhir
6. duration (numerical) : durasi kontak terakhir (dalam detik)
7. campaign (numerical) : jumlah kontak saat promosi
8. pdays (numerical) : jumlah hari yang terlewati sejak kontak terakhir
9. previous (numerical) : jumlah kontak terakhir sebelum campaign ini
10. poutcome (Categorical) : Hasil dari marketing sebelumnya
11. Feature Tujuan :
12. y (categorical) : (y/n) output dari campaign ini, feature ini adalah tujuan dari analisis yang berisi data apakah pelanggan akan berlangganan deposito berjangka atau tidak.
13. Analisis distribusi data

untuk mempermudah analisis data, feature yang ada di dalam dataset dikelompokkan menjadi 3 bagian berdasarkan tipe data, yaitu: numerik, category, dan timestamps. untuk lebih jelasnya, penjelasan mengenai 3 kelompok data tersebut adalah sebagai berikut:

1. Data numerik



Dalam feature age, dapat dilihat bahwa usia pelanggan berkisar antara 18 hingga 95 tahun, dengan usia rata-rata 40-41 tahun. distribusi usia menunjukkan variasi yang cukup besar, dengan deviasi standar 10 tahun, artinya ada penyebaran usia yang cukup lebar diantara pelanggan.

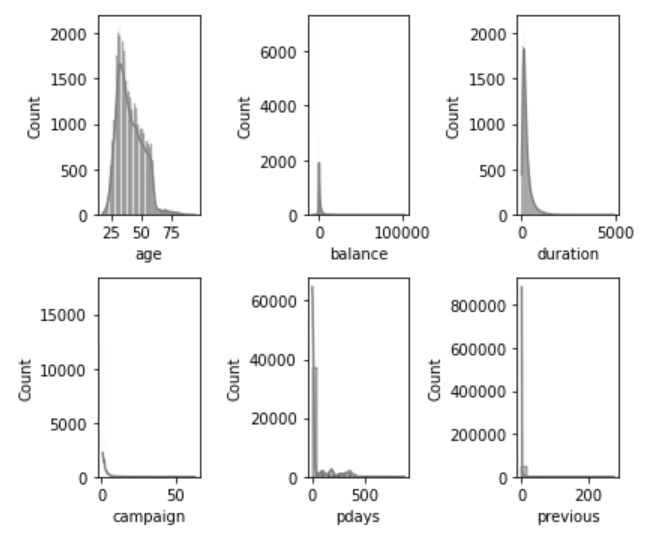
Dalam feature balance, ini memiliki nilai minimum yang sangat rendah, yaitu -8019 euro, menunjukkan adanya saldo negatif yang signifikan pada beberapa pelanggan. rata-rata saldo adalah 1.362 euro, tetapi penyebarannya sangat lebar dengan deviasi standari 2.044 euro dan nilai maksimum mencapai lebih dari 102.000 euro.

Durasi kontak telepon bervariasi dari 0 detik hingga 4,918 detik, dengan rata-rata 258 detik. Ini menunjukkan bahwa beberapa panggilan berlangsung sangat lama, sementara ada panggilan yang sangat singkat atau bahkan tidak terjadi sama sekali. Deviasi standar yang tinggi (257 detik) menunjukkan variasi besar dalam durasi panggilan.

Jumlah kontak berkisar antara 1 hingga 63 kali, dengan rata-rata sekitar 2.76 kali. Mayoritas pelanggan dihubungi 1 hingga 3 kali, seperti yang ditunjukkan oleh persentil ke-75 (3 kali). Namun, ada beberapa pelanggan yang dihubungi sangat sering selama kampanye ini.

Nilai minimum untuk pdays adalah -1, yang berarti pelanggan tidak pernah dihubungi sebelumnya. Nilai rata-rata 40.19 dan deviasi standar 100 menunjukkan bahwa ada beberapa pelanggan yang dihubungi kembali setelah jangka waktu yang cukup lama (nilai maksimum adalah 871 hari).

Banyak pelanggan tidak dihubungi sebelumnya (previous = 0 untuk kuartil ke-75). Namun, nilai maksimum menunjukkan bahwa ada pelanggan yang telah dihubungi sebanyak 275 kali dalam kampanye-kampanye sebelumnya.



Dalam grafik persebaran data numerik tersebut, diketahui bahwa sebagian besar data berdistribusi condong kanan atau *positive skew*. Artinya, banyak data numerik memiliki nilai yang rendah/kecil.

Distribusi usia pelanggan menunjukkan bahwa sebagian besar pelanggan berada dalam rentang usia antara 25 hingga 60 tahun, dengan puncak sekitar usia 30 hingga 40 tahun. Distribusi ini terlihat cenderung normal dengan penurunan jumlah pelanggan seiring bertambahnya usia.

Distribusi saldo tahunan sangat tidak merata, dengan sebagian besar pelanggan memiliki saldo yang relatif rendah, sementara ada beberapa outlier yang memiliki saldo sangat tinggi, lebih dari 100.000 euro. Ini menunjukkan ketimpangan besar dalam distribusi kekayaan pelanggan.

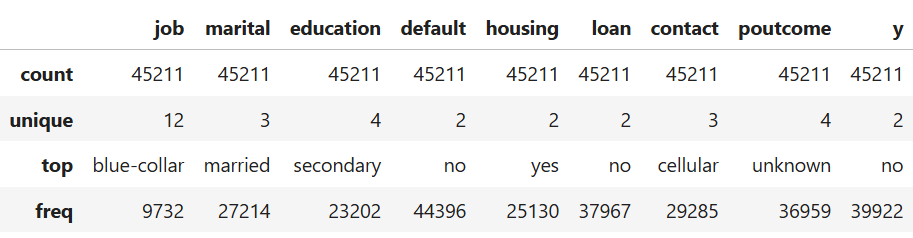
Durasi panggilan telepon bervariasi, dengan mayoritas panggilan berlangsung dalam waktu singkat (kurang dari 500 detik). Namun, ada sedikit jumlah panggilan yang berlangsung sangat lama, hingga lebih dari 4000 detik.

Sebagian besar pelanggan hanya dihubungi 1 hingga 3 kali selama kampanye, dengan hanya sedikit pelanggan yang dihubungi lebih dari 10 kali. Hal ini menunjukkan bahwa kampanye telepon tidak dilakukan secara intensif pada sebagian besar pelanggan.

Sebagian besar pelanggan belum pernah dihubungi sebelumnya (pdays = -1), dan hanya sebagian kecil yang dihubungi kembali setelah beberapa ratus hari dari kampanye sebelumnya.

Jumlah kontak sebelumnya untuk sebagian besar pelanggan adalah 0, menunjukkan bahwa sebagian besar pelanggan ini baru dihubungi dalam kampanye ini, dengan hanya beberapa pelanggan yang pernah dihubungi beberapa kali dalam kampanye sebelumnya.

1. Data Kategorik



Dalam data kategorik, ada feature yang memiliki value berupa yes-no, seperti: ‘default’, ‘housing’, ‘loan’, dan termasuk feature tujuan yaitu feature ‘y’.

dalam feature job, terdapat 12 jenis pekerjaan yang berbeda di antara pelanggan. Pekerjaan yang paling umum adalah "blue-collar" dengan 9.732 pelanggan memiliki pekerjaan ini. dalam marital, ada 3 kategori status pernikahan, dengan mayoritas pelanggan berstatus "married" (27.214 pelanggan).

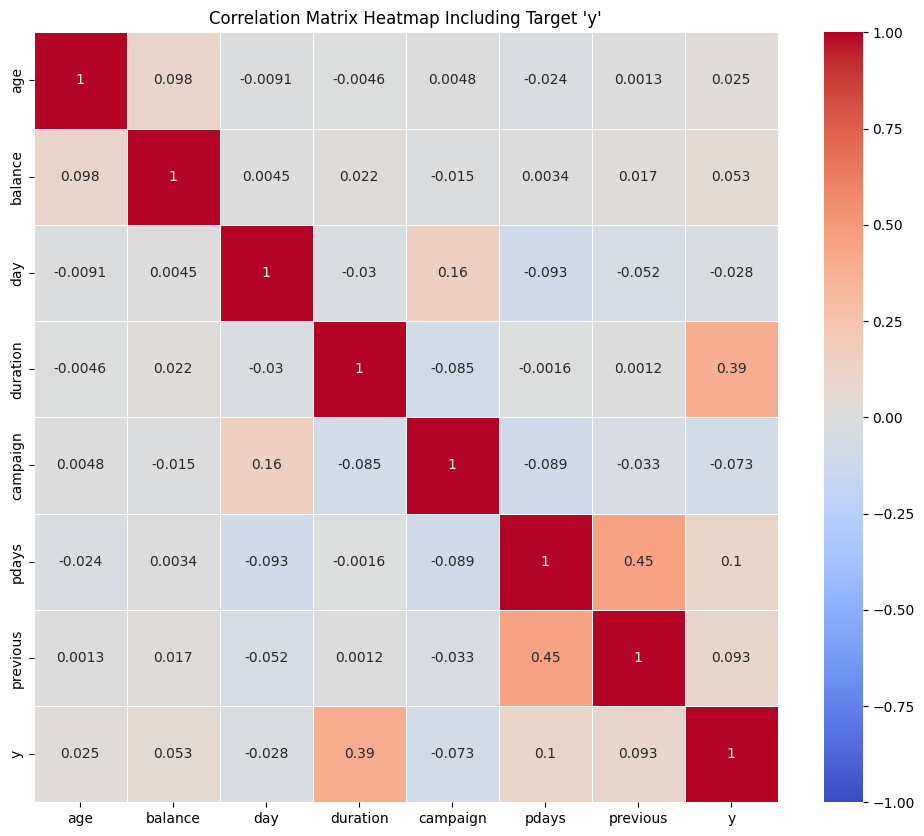
sementara dalam feature education, pendidikan terdiri dari 4 kategori, dengan kategori "secondary" (pendidikan menengah) menjadi yang paling umum, mencakup 23.202 pelanggan.

Pada data default nasabah, menunjukkan apakah pelanggan memiliki kredit macet atau tidak, dengan mayoritas besar (44.396 pelanggan) tidak memiliki kredit macet. sementara untuk kredit rumah pada housing, Sebanyak 25.130 pelanggan memiliki pinjaman rumah, sementara sisanya tidak. Mayoritas pelanggan (37.967) tidak memiliki pinjaman pribadi, hal ini tercermin dalam feature loan.

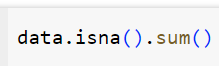
Dalam campaign sebelumnya, sebagian besar kontak dilakukan melalui "cellular" (29.285 kontak), dibandingkan dengan jenis kontak lainnya. Hasil dari kampanye telepon sebelumnya terbagi menjadi 4 kategori, dengan mayoritas hasilnya "unknown" (36.959).

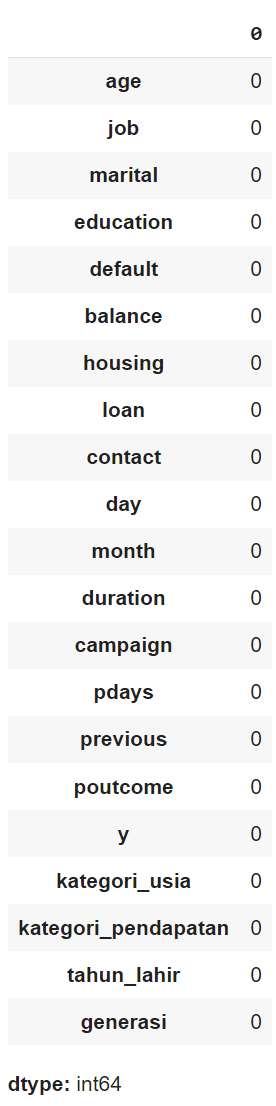
berdasarkan hasil dari variabel target yang menunjukkan apakah pelanggan berlangganan deposito berjangka atau tidak, dengan mayoritas besar pelanggan (39.922) tidak berlangganan.

1. Analisis target variable (y)



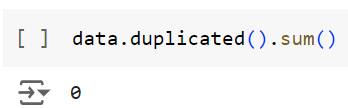
1. Preprocessing
2. Data Cleansing
3. Handle Missing values





Dari hasil analisa, dataset ini tidak memiliki missing values

1. Handle Data duplicates



Dari hasil analisa, dataset ini tidak memiliki data duplikat

1. Handle Outliers



Dari analisa boxplot di atas, sebagian besar variabel memiliki distribusi yang menunjukkan adanya banyak outlier, terutama pada variabel balance, duration, campaign, pdays, dan previous.

Dari plot histogram di atas, terlihat bahwa sebagian besar variabel memiliki distribusi yang cenderung miring ke kanan dengan banyak nilai yang terkonsentrasi pada nilai yang lebih rendah dan beberapa outlier yang memiliki nilai yang sangat tinggi.

Hasil analisa violin plot menunjukkan distribusi yang miring ke kanan untuk sebagian besar variabel, dengan kepadatan terbesar pada nilai yang lebih rendah dan adanya ekor panjang yang menunjukkan outlier.

Dalam proyek ini, kami mengeliminasi outlier menggunakan 2 cara :

1. IQR : ‘age’, ‘balance’, ‘duration’, ‘campaign’
2. Z-score : ‘pdays’, ‘previous’

Sehingga distribusi data yang dihasilkan setelah dilakukan handle outlier adalah sebagai berikut:

* 1. Normalization

Dari hasil EDA yang dilakukan, dapat disimpulkan bahwa sebagian besar data memiliki distribusi yang cenderung right skewed. Sehingga perlu dilakukan normalisasi.

Pada proyek kali ini, kami menggunakan metode logarithmic transformation untuk mengubah distribusi data menjadi mendekati normal. Distribusi data setelah dilakukan logarithmic transformation adalah sebagai berikut:

Hasil dari transformasi log ini menunjukkan bahwa beberapa variabel seperti age, duration, dan balance menunjukkan distribusi yang lebih mendekati normal, sementara variabel lain seperti campaign, pdays, dan previous masih menunjukkan distribusi yang kompleks atau sangat terfokus pada satu nilai tertentu.

* 1. Feature Encoding

Data-data kategorikal perlu diubah menjadi data numerik agar dapat di pelajari oleh model machine learning. Oleh karena itu, kami menggunakan 2 metode untuk melakukan feature encoding:

* + 1. Label Encoding

Untuk fitur-fitur yang memiliki urutan (ordinal) atau fitur boolean 🡪 education, default, housing, loan, y

* + 1. One Hot Encoding

Untuk fitur-fitur yang tidak memiliki urutan (non-ordinal)

🡪 job, marital, contact, month, poutcome

* 1. Handle Class Imbalance

Dari hasil analisa data label target (‘y’), terlihat bahwa terdapat ketimpangan besar antara data yang telah membuka deposito (y=1) dengan data yang belum membuka deposito (y=0). Dimana y=1 hanya ada 1,936 data, sedangkan y=0 terdapat 23,008 data.

Sehingga untuk dapat menghasilkan data training & test yang seimbang untuk model machine learning, kami menggunakan metode oversampling sehingga data y=0 dan y=1 seimbang 50:50 (23,008 data, baik untuk y=0 maupun y=1)

1. Feature Engineering
   1. Feature Extraction
      1. Kategori Usia

Penambahan feature kategori usia dianggap penting untuk melihat peluang calon nasabah yang membuka deposito dari Kelas Usia mereka

* + 1. Kategori Pendapatan

Penambahan feature kategori pendapatan calon nasabah yang dilihat dari balance rata rata tahunan dianggap penting untuk melihat peluang calon nasabah yang membuka deposito dari Kelas ekonomi yang tercermin pada 'balance' mereka

* + 1. Kategori Generasi

Penambahan feature kategori generasi tidak jauh berbeda dengan kategori usia, namun usia dalam hal ini dibagi berdasarkan generasi tahun kelahiran yang cenderung memiliki karakteristik yang unik.

* + 1. Durasi Kontak

Penambahan feature durasi kontak dalam satuan menit cukup penting untuk menambah insight terkait trend durasi kontak yang selama ini dilakukan dan kaitannya terhadap pembukaan deposit

* 1. Feature Selection

Dari keseluruhan fitur yang ada pada dataset, kami menggunakan semua fitur kecuali fitur ‘duration’. Hal ini disebabkan pada saat melakukan modelling menggunakan LGBM, hasil SHAP value menjadi agak bias apabila menggunakan fitur ‘duration’

1. Modelling and Machine Leaning
2. Modelling
3. Hyperparameter Tuning
4. Evaluation
5. Business Recommendation