Telemarketing dengan *Machine Learning*

Telemarketing merupakan salah satu cara yang digunakan oleh bank untuk memasarkan produknya. Kegiatan marketing ini dilakukan dengan menggunakan telepon serta *call center* untuk menjual produk pada klien. *Telemarketing* dianggap memiliki banyak keuntungan di antaranya adalah hemat biaya dan memberikan kepuasan bagi pelanggan. Tingkat keberhasilan dari *telemarketing* dalam menjual produk dapat diketahui dengan menggunakan *machine learning*.

Dengan menggunakan metode CRISP-DM, yaitu salah satu model proses untuk melakukan data mining, dapat diselesaikan *case* untuk mengetahui tingkat keberhasilan dari *telemarketing.* Tahapan dari metode CRISP-DM adalah sebagai berikut:

- ***Business Understanding***

  Di tahap ini, dibutuhkan pemahaman terkait masalah yang akan diselesaikan menggunakan metode ini. Dalam kasus ini, kita ingin mengetahui apakah klien akan berlangganan deposito berjangka setelah dilakukan pemasaran dengan *telemarketing.*

- ***Data Understanding***

  Ketika sudah memahami permasalahan yang akan diselesaikan, langkah selanjutnya adalah memahami data yang akan kita gunakan. Pada kali ini, digunakan data yang diambil dari UCI Machine Learning Repository. Data yang dipilih adalah **"bank-additional-full.csv"** dengan 41.118 *data points*, 20 variabel, di mana 10 variabelnya bersifat numerik dan sisanya adalah kategorikal. Berikut merupakan variabel yang digunakan:

  - Age (numerik)
  - Job: Jenis pekerjaan (kategorikal).
  - Marital: Status pernikahan (kategorikal)
  - Default: Apakah memiliki kredit macet? (kategorikal)
  - Housing: Apakah memiliki pinjaman rumah? (kategorikal)
  - Loan: Apakah memiliki utang pribadi? (kategorikal)
  - Contact: Jenis komunikasi (kategorikal)
  - Month: Kontak terakhir dalam setahun (kategorikal)
  - Day of week: Kontak terakhir dalam seminggu (kategorikal)
  - Duration: Durasi dari kontak terakhir (numerik)
  - Campaign: Jumlah kontak yang dilakukan selama kampanye (numerik)

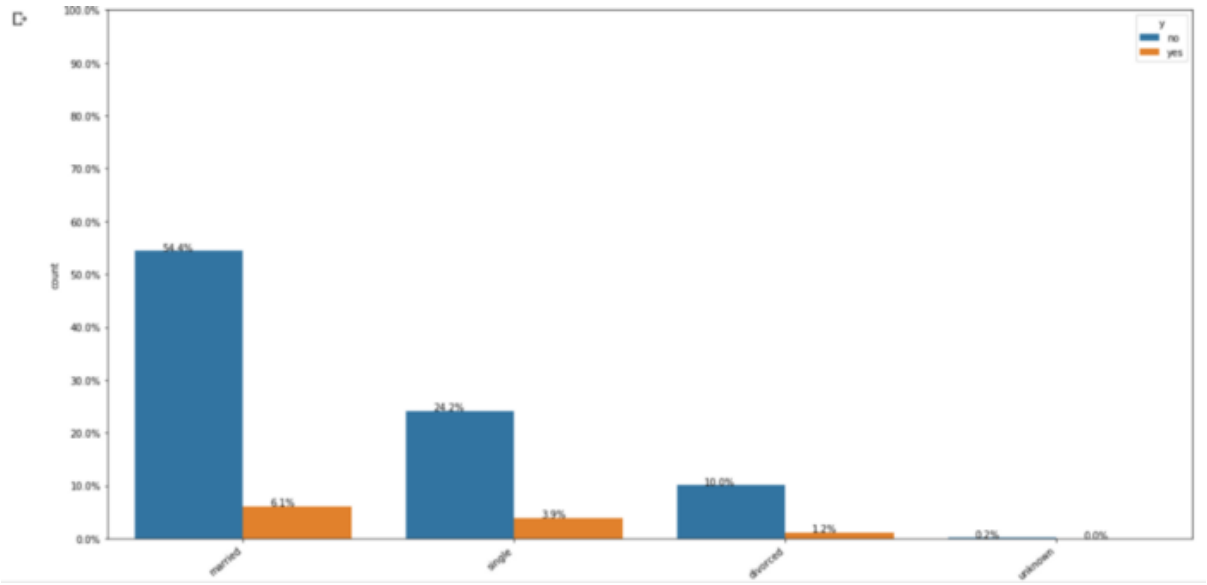
- Pdays: Jumlah hari yang telah berlalu setelah klien terakhir kali dihubungi (numerik)
- Previous: Jumlah kontak yang dilakukan sebelum ini (numerik)
- Poutcome: Hasil dari marketing sebelumnya (kategorikal)
- Emp.var.rate: Tingkat variasi pekerjaan – triwulan (numerik)
- Cons.price.idx: Indeks harga konsumen – bulanan (numerik)
- Cons.conf.idx: Indeks kepercayaan konsumen – bulanan (numerik)
- Euribor3m: Kurs euribor 3 bulan – harian (numerik)
- Nr. Employed: jumlah karyawan – triwulan (numerik)

➢ ***Data Preparation***

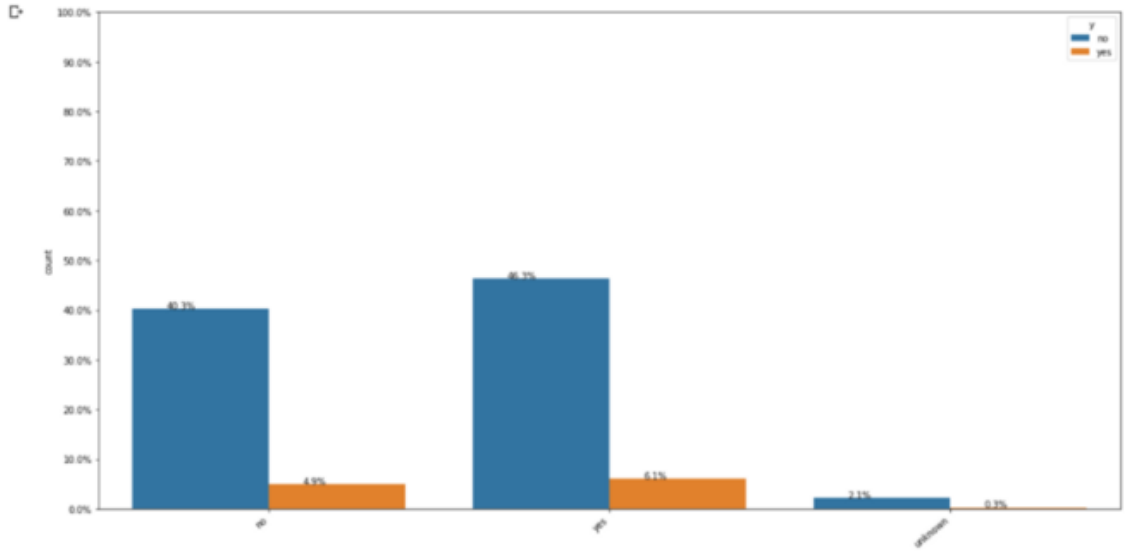
Setelah menetapkan variabel apa saja yang akan digunakan, selanjutnya kita akan melakukan *data preparation*. Dikarenakan dataset yang kita miliki terdapat 10 variabel yang bentuknya kategorikal, maka perlu untuk dilakukan *encode* ke bentuk numerik agar dapat diterapkan model *machine learning.* Terdapat dua skema, namun skema yang paling populer adalah **one hot encoding** dengan membuat kolom baru yang bernilai biner (0 atau 1). Selain melakukan *encode,* proses dari tahapan ini adalah dengan mengatasi *missing values.* Namun, di dalam dataset yang kita gunakan tidak terdapat *missing values*. Jika terdapat, maka hal yang perlu dilakukan hanya menghapusnya atau menggunakan imputasi untuk menanganinya.

Lalu, kita juga perlu menangani data yang terduplikasi dengan menghapusnya. Langkah yang dilakukan selanjutnya adalah memisahkan antara dependen dengan independen variabel. Hal ini diperlukan sebelum lanjut ke tahap *modelling.* Terakhir, sangat penting untuk membagi dataset yang kita miliki menjadi dataset *train, test,* dan CV. Jika tidak, maka akan terjadi kebocoran data. Pada kali ini, kita akan menerapkan masing-masing 64%, 16%, dan 20% untuk dataset *train, test,* dan CV.

Kita juga melakukan *Explanatory Data Analysis (*EDA) untuk melihat apakah ada pola dalam data yang kita miliki. Kita akan melakukan beberapa analisis univariat untuk mengetahui variabel mana yang tidak terlalu penting. Terdapat beberapa contoh analisis univariat dari variabel kategorikal dan numerik:

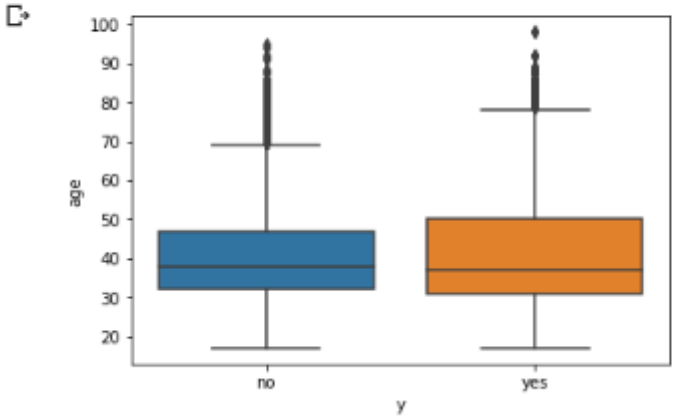


Marital status



Housing Loan

```
[ ]  %matplotlib inline
     sns.boxplot(data=data, x="y", y="age")
     plt.show()
```



Age

Dari grafik di atas, diketahui bahwa kebanyakan klien berstatus menikah dan memiliki pinjaman rumah. Klien yang memiliki atau tidak memiliki deposito berjangka memiliki usia rata-rata 38-40 tahun.

- ***Modelling***

  Kita akan membuat model sederhana dengan regresi logistik dengan menggunakan variabel *duration* untuk melihat bagaimana model dengan variabel ini. Hasilnya adalah sangat bagus. Namun, kita tidak dapat menggunakan variabel ini ke dalam model karena sangat berkorelasi dengan variabel dependen dan memasukkan variabel ini ke dalam model menjadikan modelnya tidak masuk akal. Sehingga, kita harus menghapusnya. Setelah menghapusnya, kita melakukan regresi logistik tanpa variabel *duration*. Setelah itu dilakukan *linear SVM, random forest,* dan *XGBoost.*

- ***Evaluation***

  Evaluasi dilakukan untuk melihat apakah hasil dari model sudah memenuhi tujuan dari bisnis atau belum. Ketika dirasa belum, maka akan dilakukan tahapan-tahapan sebelumnya sehingga ditemukan hasil yang terbaik. Hasil dari kasus ini, ditemukan bahwa *XGBoost* memberikan skor tertinggi pada tes AUC dengan nilai 0,803. Dapat diketahui juga bahwa variabel yang paling penting adalah variabel numerik. Fitur yang paling penting dalam memprediksi apakah klien akan membeli deposito berjangka adalah:

  - Nr. Employed
  - Emp.var.rate
  - Poutcome_success
  - Euribor3m

- ***Deployment***

  Setelah dilakukan tahapan-tahapan untuk menentukan model mana yang paling cocok dan variabel apa sajakah yang berperan penting, maka dapat dilakukan prediksi dari tingkat keberhasilan dari *telemarketing.*

Dengan menerapkan algoritma logistik dan regresi, model berhasil dibangun. Dengan model yang ada, bank dapat memprediksi respon klien terhadap *telemarketing* sebelum melakukan panggilan terhadap klien. Bank juga dapat mengutamakan klien yang telah diklasifikasikan sebagai individu yang sangat mungkin untuk membeli deposito berjangka.

References:

Roy, Sukanta. 2019. “Machine Learning Case Study: A data-driven approach to predict the success of bank telemarketing.” Machine Learning Case Study: A data-driven approach to predict the success of bank telemarketing | by Sukanta Roy | Towards Data Science