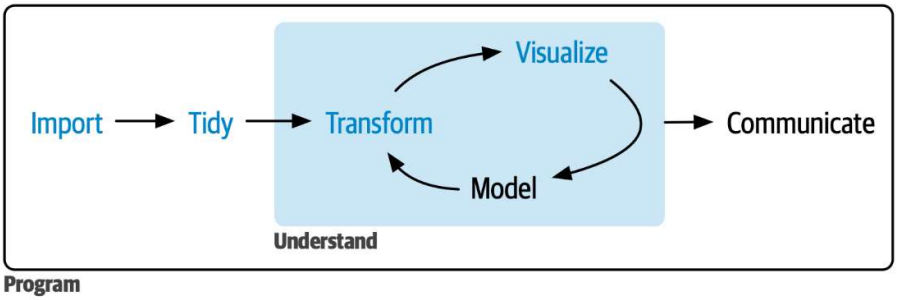
Exercise 1

Saya menyukai ilustrasi tersebut dari buku R for Data Science [1].



Hal pertama yang seorang data analyst hadapi ialah meng-import data, atau data scraping. Data dapat diambil dari database perusahaan, online repository, maupun experimen baru seperti survey atau experimental setup. Namun, data yang di-scrape pasti berupa data kotor. Setelah meng-import, kita akan meng-tidy atau membersihkan data. Tidy biasanya mencangkup formatting, menghilangkan atau meng-impute data yang hilang.

Setelah mendapatkan data yang siap untuk dianalisa, maka biasanya cycle ini akan masuk ke dalam repetitive proses untuk mencapai target yang dimiliki. Biasanya dimulai oleh proses transformasi data. Data transformasi bisa berupa banyak hal, tergantung jenis data yang sedang di proses. Contoh, dalam signal processing signal yang ada di time domain dapat ditransform menjadi frequency domain dengan Fourier analysis. Kemudian kita harus melakukan analisa berupa visualisasi. Visualisasi diperlukan untuk menjamin supaya data kita memiliki sifat atau fitur statistik yang proper untuk machine learning. Setelah itu, kita akan men-fit data kita ke model, dan analisa performa model tersebut. Biasanya, model pertama tidak mencapai performa optimal, sehingga diperlukan transformasi, visualisasi, dan modelling cycle yang berulang sampai mencapai target dengan teknik-teknik yang berbeda.

Setelah menemukan model yang optimal, solusi untuk masalah yang ada, maka hasilnya harus di komunikasikan kepada stakeholder. Komunikasi sangatlah penting karena tanpa komunikasi yang efektif, maka temuan dan solusi dari proses ini akan sia-sia. Penting untuk memahami target audience kita, apakah mereka dari bidang bisnis atau IT, sehingga media presentasi dan bahasa yang digunakan menjadi efektif.

Contoh proyek adalah membuat AI yang harus memprediksi tingkat atau probabilitas nasabah gagal membayar kredit. Hal pertama ialah mengimpor data nasabah dan rekor history dari database bank. Kemudian, data tersebut harus di format (tidy), karena kemungkinan besar rekor-rekor tersebut berasal dari berbagai macam database, format, dan struktur, sehingga data dari berbagai database dapat hidup di dalam struktur yang sama untuk memulai analisa data. Missing values dapat di interpolitasi, tergantung jumlah dan kualitas data yang tersedia.

Kemudian data harus ditransformasikan supaya secara statistik dapat digunakan untuk modelling. Untuk data-data kategorikal nasabah seperti jenis kelamin, provinsi KTP, dan status pernikahan bisa dirubah menjadi continuous seperti teknik one-hot-encode. Data harus di normalisasikan atau scale supaya tidak ada bias di dalam data. Data seperti jumlah kredit akan skewed ke kanan, sehingga wajib di normalisasikan. Visualisasi penting untuk memberikan

insight terhadap data, seperti memberikan gambaran apabila nasabah yang sering terlambat membayar kredit lebih cenderung untuk gagal membayar kredit sama sekali. Kemudian, saya akan memulai dengan model-model sederhana sebagai patokan awal untuk data yang telah disiapkan, seperti Logistic Regression untuk binary classification terkait kredit gagal. Apabila hasil kurang memuaskan, kita dapat mentransformasi ulang data untuk model yang lebih modern seperti transformer based Huggingface models atau boosting algorithm seperti Catboost.

Ketika hasil sudah memuaskan, maka kita akan melakukan presentasi terhadap bisnis. Apabila model memiliki hasil yang memuaskan, maka kita dapat mempresentasikan hasil dengan jargon-jargon yang digunakan bisnis. Dengan model ini, kita dapat mendeteksi dini kredit gagal dengan akurasi sekian persen, sehingga banyak kredit gagal dapat dicegah dan membuat bisnis hemat sekian banyak IDR. Sebagai call of action, model ini sebaiknya dideploy untuk analisa kredit baru dan kredit yang sedang aktif.