BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 7

UKURAN KEMIRIPAN (SIMILARITAS)

LEARNING OUTCOMES

- Setelah mempelajari materi ini peserta kuliah diharapkan mampu mengidentifikasikan dan memahami:
  - Calculating similarity of objects described by data; Distance metrics for calculating similarity.
  - Searching for similar entities;
  - Using similarity for prediction;
  - Clustering as similarity-based segmentation.Nearest neighbor methods;
  - Various clustering methods;

OUTLINE MATERI :

1. Ukuran Similaritas Dan Konsep Jarak
2. Konsep Dasar Nearest-Neighbor (NN)
3. Kesimpulan

PENDAHULUAN

Ukuran kesamaan mendasari banyak metode dalam data sains dan solusi untuk masalah bisnis. Jika dua entitas (orang, perusahaan, produk) serupa dalam beberapa hal, mereka sering serupa dalam karakteristik lain. Prosedur data sains sering didasarkan pada pengelompokan berbagai hal berdasarkan kesamaan atau mencari kesamaan "yang tepat". Secara implisit konsep ini ada dalam Sesi-Sesi sebelumnya di mana prosedur pemodelan membuat batas-batas untuk mengelompokkan sampel berdasarkan kesamaan untuk variabel target. Dalam Sesi ini kita akan melihat kesamaan secara langsung, dan menunjukkan bagaimana hal itu berlaku untuk berbagai tugas yang berbeda. Berbagai macam kasus bisnis melibatkan penalaran dari sampel sampel serupa:

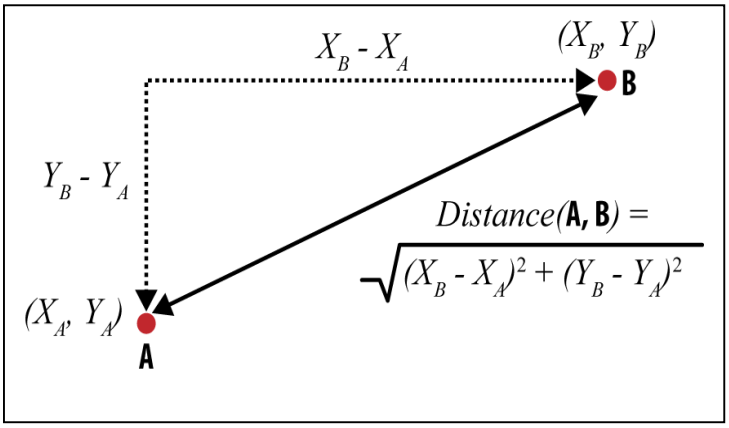
Similaritas Dan Jarak

Jika sebuah objek dapat direpresentasikan sebagai data, dapat diperkenalkan konsep tentang kesamaan antara objek, atau jarak antara objek. Misalnya, representasi data yang telah digunakan di sepanjang Kuliah sejauh ini: mewakili setiap objek sebagai vektor fitur. Kemudian, semakin dekat dua objek dalam ruang yang ditentukan oleh fitur, maka objek objek itu akan semakin mirip.

Ketika kita membangun dan menerapkan model prediktif, tujuannya adalah untuk menentukan nilai karakteristik target. Dengan demikian, kita telah menggunakan kesamaan implisit dari objek. Banyak metode dalam data sains dapat perlakukan dengan cara ini: sebagai metode untuk mengatur sampel data (representasi objek penting) sehingga sampel yang berdekatan satu sama lain diperlakukan sama untuk beberapa tujuan. Kedua pohon klasifikasi dan pengklasifikasi linier menetapkan batas antar daerah dari klasifikasi yang berbeda. Mereka memiliki pandangan yang sama bahwa sampel dalam wilayah yang sama harus serupa; apa yang berbeda di antara metode adalah bagaimana wilayah dibagi dan ditemukan. Dibutuhkan metode dasar untuk mengukur kesamaan atau jarak. Apa artinya dua perusahaan atau dua konsumen itu serupa? Perhatikan dua contoh dari domain aplikasi kartu kredit yang disederhanakan:

Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1

Item-item data ini memiliki beberapa atribut, dan tidak ada metode tunggal terbaik untuk menguranginya ke satu kesamaan atau pengukuran jarak. Ada banyak cara untuk mengukur kesamaan atau jarak antara A dan B. Permulaan yang baik adalah dengan mengukur jarak menggunakan geometri dasar.



Gambar 6-1. Euclidean distance.

Ukuran ini disebut jarak Euclidean antara dua titik. Jarak Euclidean tidak terbatas pada dua dimensi. Jika A dan B adalah objek yang dijelaskan oleh tiga fitur, mereka dapat diwakili oleh titik-titik dalam ruang tiga dimensi dan posisi mereka kemudian akan direpresentasikan sebagai (xA, yA, zA) dan (xB, yB, zB). Jarak antara A dan B kemudian akan mencakup (zA-zB)2. Kita dapat menambahkan banyak fitur secara acak, masing-masing dimensi baru. Ketika suatu objek digambarkan oleh n fitur, n dimensi (d1, d2,…, dn), persamaan umum untuk jarak Euclidean dalam n dimensi ditunjukkan di bawah ini:

$$\sqrt{(d1, A - d1, B)2 + (d2, A - d2, B)2 + \ldots + (dn, A - dn, B)2}$$

Kini kita memiliki metrik untuk mengukur jarak antara dua objek yang dijelaskan oleh vektor fitur numerik — rumus sederhana berdasarkan jarak fitur individual objek.

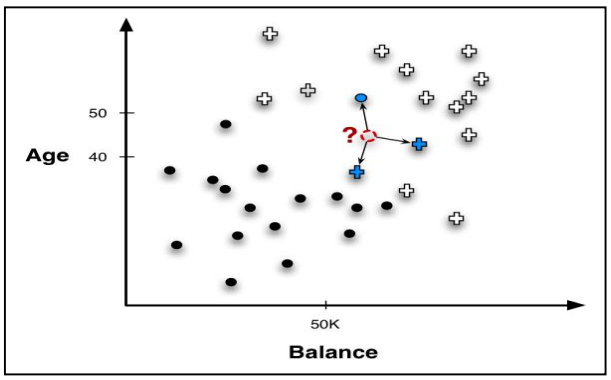
Jarak ini hanyalah angka — ia tidak memiliki unit, dan tidak ada interpretasi yang berarti. Ini hanya sangat berguna untuk membandingkan kemiripan satu pasang contoh dengan pasangan lain. Ternyata membandingkan persamaan sangat berguna.

Konsep Dasar Nearest-Neighbor (NN)

Setelah kita memiliki cara untuk mengukur jarak, kita dapat menggunakannya untuk berbagai tugas analisis data yang berbeda. Perhatikan salah satu contoh dari awal Sesi ini, kita bisa menggunakan ukuran ini untuk menemukan perusahaan yang paling mirip dengan pelanggan korporat, atau konsumen online yang paling mirip dengan pelanggan ritel yang dimiliki. Setelah kita menemukan ini, dapat diambil tindakan yang diperlukan sesuai konteks bisnis. Untuk pelanggan korporat, IBM melakukan ini untuk membantu mengarahkan para sales. Pengiklan online melakukan ini untuk menentukan target iklan. Instans yang paling mirip disebut tetangga terdekat (nearest neighbors/ NN).

Nearest Neighbors (NN) Dalam Model Prediktif

Konsep NN dapat digunakan untuk melakukan pemodelan prediktif dengan cara yang berbeda. Ingatlah semua yang Anda ketahui tentang pemodelan prediktif dari sesi sesi sebelumnya. Untuk menggunakan kemiripan dalam pemodelan prediktif, prosedur dasarnya sangat sederhana: diberi contoh baru yang variabel targetnya ingin kita prediksi, kita pindai melalui semua contoh pelatihan dan pilih beberapa yang paling mirip dengan contoh baru.



Gambar 6-2. NN classification. Titik yang akan diberi label diberi tanda tanya dan menurut gambar di atas akan memperoleh label **+** *karena mayoritas tetangganya berlabel* **+**.

Kelas instans baru akan diprediksi berdasarkan nilai target (target) tetangga terdekatnya. Bagaimana melakukan langkah terakhir itu perlu didefinisikan; untuk saat ini, katakan saja bahwa kita memiliki beberapa fungsi gabungan (seperti pemungutan suara atau rata-rata) yang beroperasi pada nilai target yang dimiliki tetangga. Fungsi penggabungan akan memberi kita jalan untuk melakukan prediksi.

Klasifikasi

Gambar 6-2 menunjukkan instans baru yang labelnya ingin kita prediksikan, ditandai dengan "?." Mengikuti prosedur dasar yang diperkenalkan di atas, tetangga terdekat (dalam contoh ini, tiga di antaranya) diambil dan variabel targetnya yang diketahui (kelas) dikonsultasikan. Dalam hal ini, dua contoh positif dan satu negatif. Apa yang seharusnya menjadi fungsi penggabungan kita? Fungsi penggabungan sederhana dalam hal ini adalah suara terbanyak, sehingga kelas yang diprediksi akan positif. Menggali lebih dalam, pikirkan masalah pemasaran kartu kredit. Tujuannya adalah untuk memprediksi apakah pelanggan baru akan menanggapi penawaran kartu kredit berdasarkan, pelanggan lain yang mirip merespons tawaran itu. Data tentang terjadinya respons ditunjukkan pada Tabel 6-1.

Tabel 6.1. Contoh penerapan Nearest Neighbor.

Pelanggan	Usia	Pendapatan ($)	Jumlh Kartu Kredit	Respons (Target)	Jarak dari David
David	37	50	2	?	0
John	35	35	3	Ya	15,16
Rachael	22	50	2	Tidak	15
Ruth	63	200	1	Tidak	152,23
Jefferson	59	170	1	Tidak	122
Norah	25	40	4	Ya	15,74

Dalam contoh ini, ada lima pelanggan yang sebelumnya telah dihubungi dengan tawaran kartu kredit. Untuk masing-masing dari mereka, disimpan nama, usia, pendapatan, jumlah kartu yang sudah mereka miliki, dan apakah mereka menanggapi tawaran itu. Untuk orang baru, David, kita ingin memprediksi apakah dia akan menanggapi tawaran itu atau tidak.

Kolom terakhir pada Tabel 6-1 menunjukkan perhitungan jarak, menggunakan Persamaan 6-1, bagaimana jarak setiap instans dari David. Tiga pelanggan (John, Rachael, dan Norah) cukup mirip dengan David, dengan jarak sekitar 15. Dua pelanggan lainnya (Ruth dan Jefferson) agak jauh. Oleh karena itu, tiga tetangga terdekat David adalah Rachael, lalu John, lalu Norah. Jawaban mereka adalah Tidak, Ya, dan Ya. Jika kita mengambil suara mayoritas dari nilai-nilai ini, kita memprediksi Ya (David akan merespon). Contoh Ini terkait beberapa masalah penting dengan menggunkan konsep NN: berapa banyak tetangga yang harus kita gunakan? Haruskah mereka memiliki bobot yang sama dalam fungsi penggabungan?

Pendugaan Probabilitas

Telah ditekankan bahwa dalam beberapa tugas bukan hanya bertujuan untuk mengklasifikasikan suatu instans baru, tetapi untuk memperkirakan kemungkinannya — memberikan skor, karena skor

memberikan lebih banyak informasi daripada hanya keputusan Ya / Tidak. Klasifikasi tetangga terdekat dapat digunakan untuk melakukan hal ini dengan mudah. Pertimbangkan lagi tugas klasifikasi untuk memutuskan apakah David akan menjadi responden atau tidak. Tetangga terdekatnya (Rachael, John, dan Norah) masing-masing memiliki kelas Tidak, Ya, dan Ya. Jika kita memberi nilai Ya = 1 dan Tidak = 0, kita dapat mengitung rata rata menjadi 2/3 untuk David. Jika kita melakukan ini dalam praktek, kita mungkin ingin menggunakan lebih dari hanya tiga tetangga terdekat untuk menghitung perkiraan probabilitas.

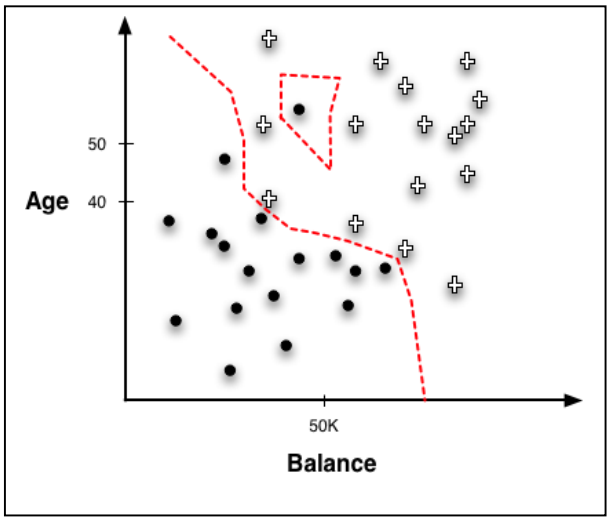
Regresi

Setelah menentukan tetangga terdekat, kita dapat menggunakannya untuk tugas penambangan prediktif manapun dengan menggabungkannya dalam beberapa cara yang berbeda. Kita baru saja melihat bagaimana melakukan klasifikasi dengan mengambil suara mayoritas dari target. Kita dapat melakukan regresi dengan cara yang serupa. Anggap kita memiliki dataset yang sama seperti pada Tabel 6-1, tetapi kali ini kita ingin memprediksi Pendapatan David. Kita tidak akan mengulangi penghitungan jarak, tetapi berasumsi bahwa tiga tetangga terdekat Daud adalah Rachael, John, dan Norah seperti sebelumnya. Pendapatan mereka masing-masing adalah 50, 35, dan 40 (dalam ribuan). Dengan nilai nilai ini kemudian digunakan untuk menghasilkan prediksi pendapatan David. Kita bisa menggunakan rata-rata (sekitar 42) atau median (40).

Penting untuk dicatat bahwa dalam menentukan tetangga, kita tidak menggunakan variabel target karena kita mencoba untuk memperkirakannya. Dengan demikian Penghasilan tidak akan masuk ke dalam perhitungan jarak seperti pada Tabel 6-1. Namun, kita bebas menggunakan variabel lain yang nilainya tersedia untuk menentukan jarak.

Interpretasi Geometric, Overfitting, dan Control Komplexitas

Seperti halnya model lain yang telah didiskusikan, adalah instruktif untuk memvisualisasikan daerah klasifikasi yang dihasilkan dengan metode tetangga terdekat. Meskipun tidak ada pembatas yang dihasilkan secara eksplisit, ada wilayah implisit yang dibuat oleh instans. Daerah ini dapat dihitung dengan menganalisa setiap titik secara sistematis dalam ruang contoh, menentukan klasifikasi setiap titik, dan membangun batas di mana klasifikasi berubah.



Gambar 6-3. Batas yang diciptakan oleh klasifier 1-NN.

Gambar 6-3 mengilustrasikan daerah yang terbentuk oleh classifier 1-NN pada instas dalam domain “Write-off” kita. Bandingkan dengan daerah-daerah pohon klasifikasi dan daerah-daerah yang dibuat oleh batas linear.

Perhatikan bahwa batas bukan garis, juga tidak ada bentuk geometris yang dapat dikenali; mereka tidak menentu dan mengikuti batas antara contoh pelatihan dari kelas yang berbeda. Pemisah tetangga terdekat mengikuti batasan yang sangat spesifik di sekitar contoh pelatihan. Perhatikan juga satu contoh negatif yang terisolasi di dalam contoh positif menciptakan "pulau negatif" di sekitarnya. Titik ini mungkin dianggap sebagai noise atau pencilan.

Beberapa kepekaan terhadap pencilan ini disebabkan oleh penggunaan penggolongan 1-NN, yang hanya mengambil satu contoh, dan begitu juga dengan batas yang lebih tidak menentu daripada yang rata-rata beberapa tetangga. Secara umum, batas-batas konsep yang tidak teratur adalah karakteristik dari semua penggolong tetangga terdekat, karena mereka tidak memaksakan bentuk geometris tertentu pada pengklasifikasi. Sebaliknya, mereka membentuk batas-batas di ruang contoh yang disesuaikan dengan data spesifik yang digunakan untuk pelatihan.

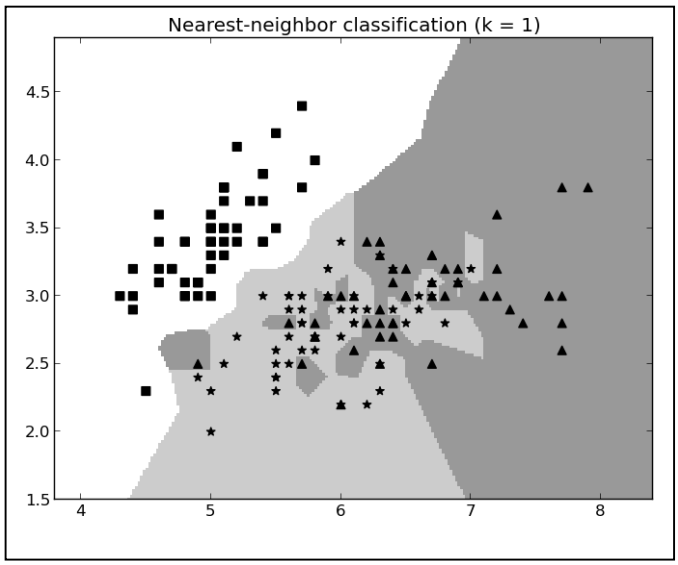
Keadaan ini memaksa kita untuk kembali ke diskusi mengenai pengendalian overfitting dan kompleksitas. Jika Anda berpikir bahwa 1-NN sangat bersifat overfitt, maka Anda benar. Bahkan, pikirkan apa yang akan terjadi jika Anda mengevaluasi penggolongan 1-NN pada data pelatihan. Ketika mengklasifikasikan setiap titik data pelatihan, setiap metrik jarak yang wajar akan mengarah pada pengambilan titik pelatihan itu sendiri sebagai tetangga terdekatnya adalah diri sendiri! Maka nilainya sendiri untuk variabel target akan digunakan untuk memprediksi dirinya, inilah klasifikasi sempurna. Hal yang sama berlaku untuk regresi. Teknik 1-NN menghafal data pelatihan. Karena tabel pencarian tidak memiliki gagasan kesamaan, itu hanya diprediksi sempurna untuk contoh pelatihan yang tepat, dan memberikan beberapa prediksi standar untuk semua yang lain. Penggolongan 1-NN memprediksi dengan sempurna untuk contoh pelatihan, tetapi juga dapat membuat prediksi yang sering masuk akal pada contoh lain: menggunakan contoh pelatihan yang paling mirip.

Jadi, dalam konteks overfitting dan penghindarannya, k dalam classifier k-NN adalah parameter kompleksitas. Pada ekstrim yang satu, kita dapat mengatur k = n dan kita tidak mengizinkan banyak kerumitan sama sekali dalam model. Seperti yang dijelaskan sebelumnya, model n-NN (mengabaikan bobot kesamaan) hanya memprediksi nilai rata-rata dalam dataset untuk setiap kasus. Pada ekstrem yang lain, kita dapat menetapkan k = 1, dan kita akan mendapatkan model yang sangat kompleks, yang menempatkan batas-batas rumit sehingga setiap contoh pelatihan akan berada di wilayah yang diberi label oleh kelasnya sendiri.

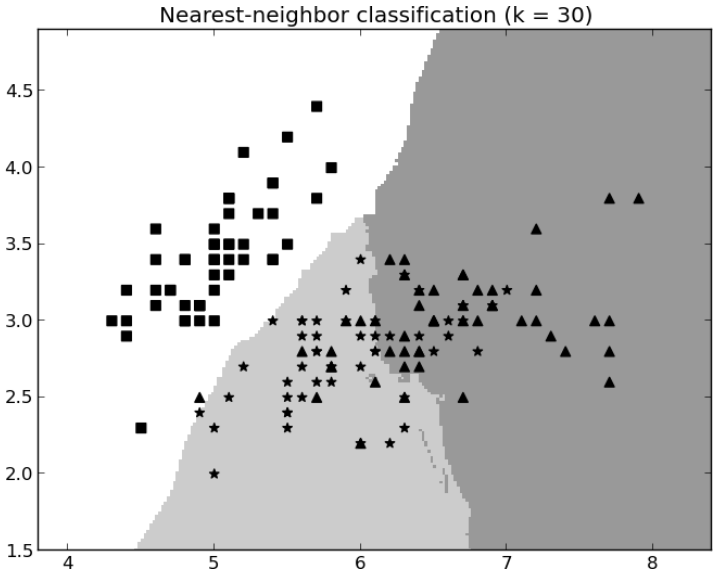
Kini kembali ke pertanyaan sebelumnya: bagaimana seharusnya seseorang memilih k? Kita dapat menggunakan prosedur yang sama yang didiskusikan dalam “Metode Umum untuk Menghindari Overfitting” untuk pengaturan parameter kompleksitas lainnya: kita dapat melakukan validasi silang atau uji coba lainnya pada data pelatihan, untuk berbagai nilai berbeda dari k, mencari salah satu yang memberikan kinerja terbaik pada data pelatihan. Kemudian ketika kita telah memilih

nilai k, kita membangun model k-NN pada seluruh data latihan. Karena prosedur ini hanya menggunakan data pelatihan, kita masih dapat mengevaluasinya pada data uji dan mendapatkan perkiraan yang tidak bias atas kinerja generalisasinya. Metode penambangan data biasanya memiliki kemampuan untuk melakukan validasi silang bersarang untuk mengatur k secara otomatis.

Gambar 6-4 dan Gambar 6-5 menunjukkan batas-batas yang berbeda yang dibuat oleh penggolong tetangga terdekat. Di sini masalah tiga kelas sederhana diklasifikasikan menggunakan jumlah tetangga yang berbeda. Pada Gambar 6-4, hanya satu tetangga yang digunakan, dan batas-batasnya tidak menentu dan sangat spesifik untuk contoh-contoh pelatihan dalam dataset. Pada Gambar 6-5, 30 tetangga terdekat dirata-ratakan untuk membentuk klasifikasi. Batas-batas jelas berbeda dari Gambar 6-4 dan jauh lebih sedikit bergerigi. Batas-batas untuk k-NN lebih ditentukan oleh data.



Gambar 6-4. Batas batas klasifikasi yang dibentuk oleh masalah klasifikasi 3 kelas menggunakan 1-NN (single nearest neighbor).



Gmbar 6-5. Batas klasifikasi yang terbentuk oleh pengklasifikasian 3 kelas menggunakan 30-NN (averaging 30 nearest neighbors).

KESIMPULAN

Konsep dasar kesamaan antara item data terjadi di seluruh Teknik penambangan data. Dalam sesi ini, pertama-tama kita membahas berbagai penggunaan kesamaan mulai dari menemukan entitas yang mirip (atau objek) berdasarkan deskripsi data, hingga pemodelan prediktif, hingga entitas pengelompokan. Kita telah membahas berbagai penggunaan dan diilustrasikan dengan contoh-contoh.

Konsep yang sangat umum untuk kesamaan dua entitas adalah jarak antara mereka dalam ruang contoh yang ditentukan oleh representasi vektor fitur mereka. Kita mempresentasikan persamaan dan perhitungan jarak, umumnya dan dalam detail teknis. Jjuga telah diperkenalkan keluarga metode, yang disebut metode tetangga terdekat (NN), yang melakukan tugas prediksi dengan menghitung secara eksplisit kesamaan antara contoh baru dan satu set contoh pelatihan (dengan nilai yang diketahui untuk target). Setelah kita menentukan tetangga terdekat (contoh yang paling mirip) kita dapat menggunakannya untuk berbagai tugas penambangan data: klasifikasi, regresi, dan pemberian skor instans. Akhirnya, telah ditunjukkan bagaimana konsep dasar yang sama — kesamaan — mendasari metode umum dalam penambangan data tanpa pengawasan: pengelompokan.

Telah pula dibahas konsep penting lain yang digunakan untuk analisis data eksplorasi lebih lanjut. Ketika mengeksplorasi data, terutama dengan metode yang tidak diawasi, kita biasanya menghabiskan lebih sedikit waktu di awal fase pemahaman bisnis dari proses penambangan data, tetapi lebih banyak waktu dalam tahap evaluasi, dan dalam iterasi di sekitar siklus.

DAFTAR PUSTAKA

1. Foster Provost & Tom Fawcett (2013) Data Science for Business: What you need to know about data mining and data analytic thinking, O'Reilly, ISBN: 978-1-449-36132-7.

2. Sharda, R., Delen, D., Turban, E., (2018). Business intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.