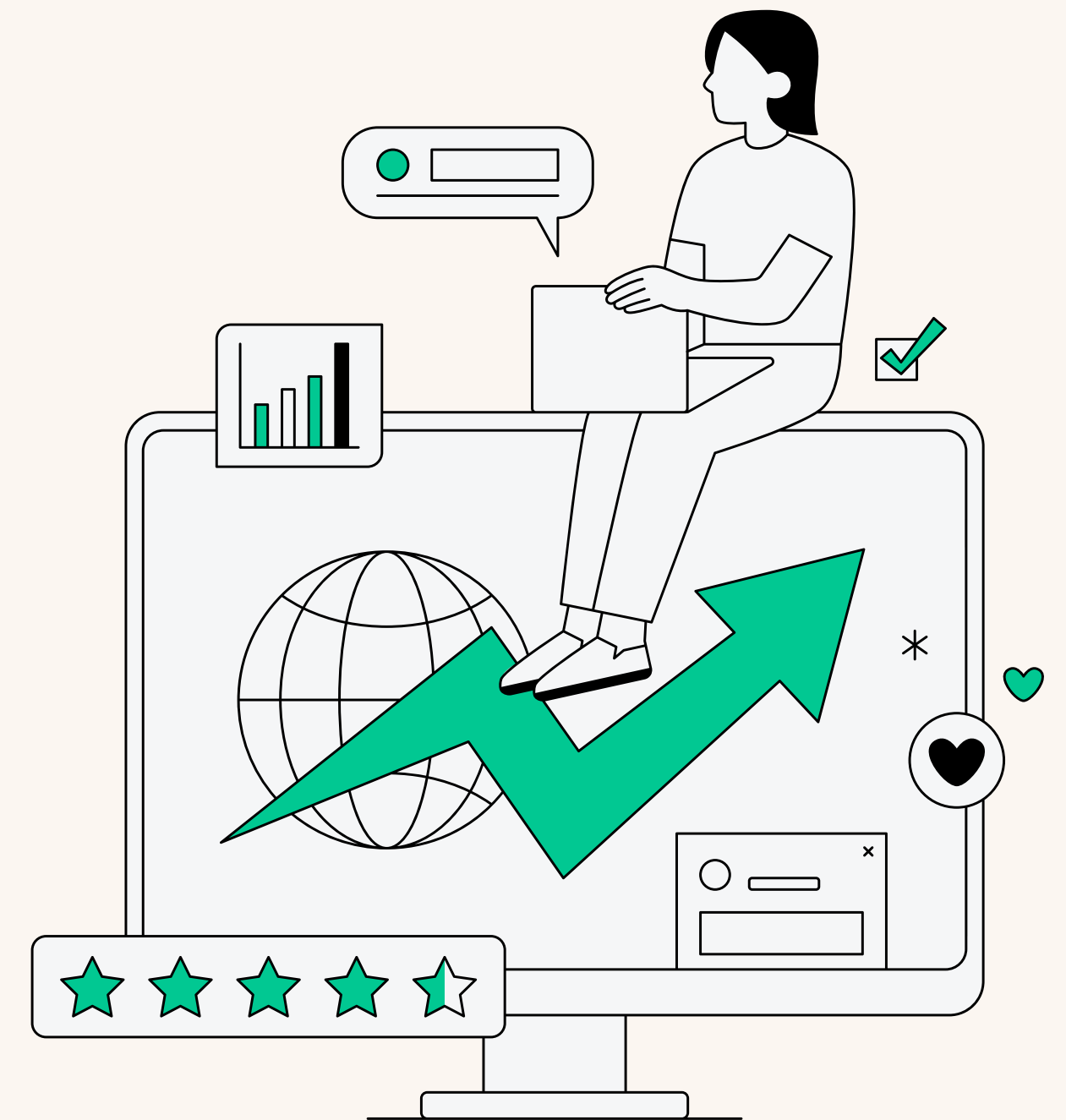


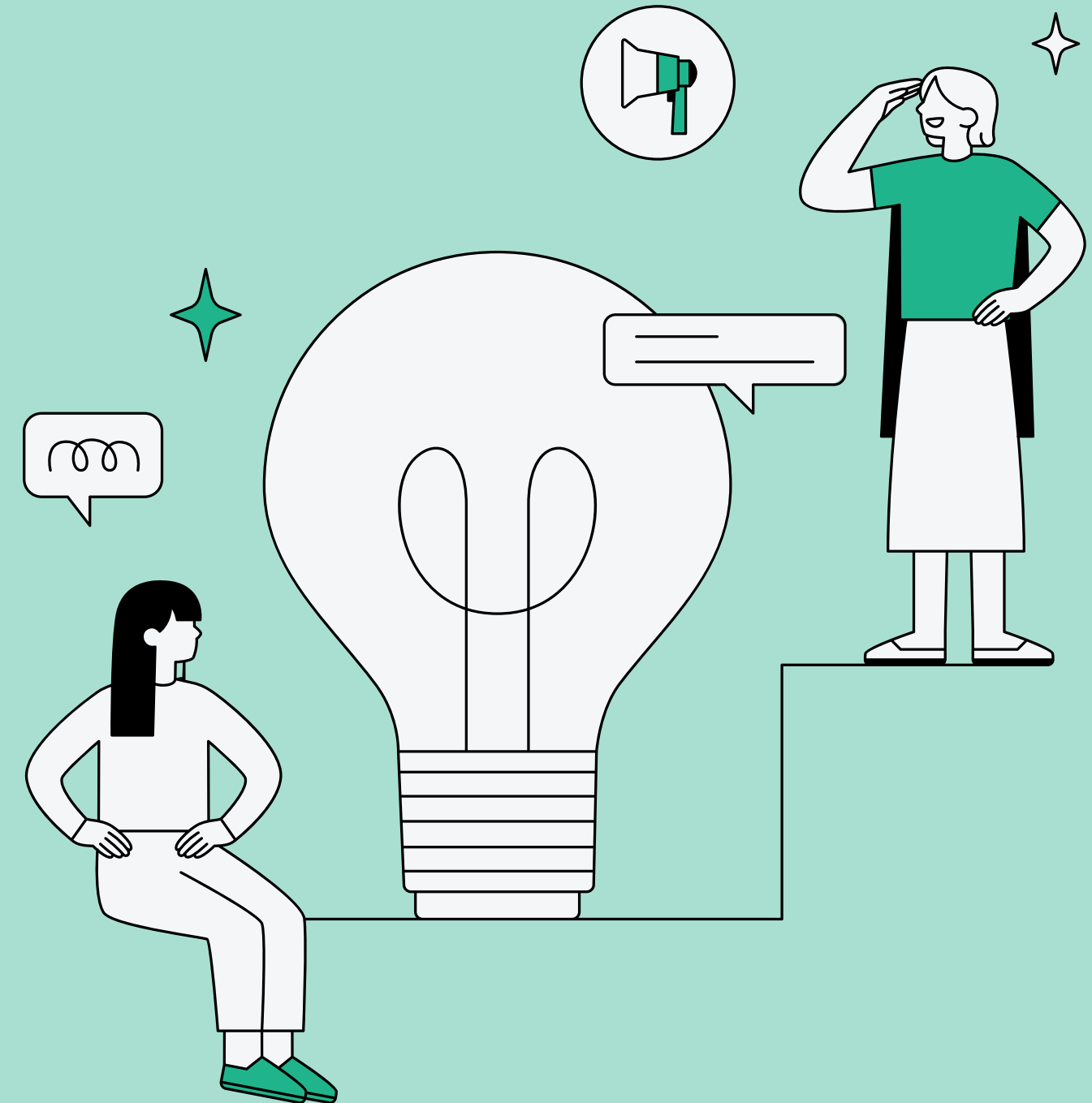
Analysis Feature Importance

Michael Habelard H - 71487




Dataset

- Sumber: SuperStore_Order.csv
- Jumlah data: ~51000
- Variabel target:
 - profit
 - shipping_cost
- Variabel independen (fitur)
 - category
 - order_priority
 - sub_category
 - ship_mode
 - market
 - region
 - country
 - state



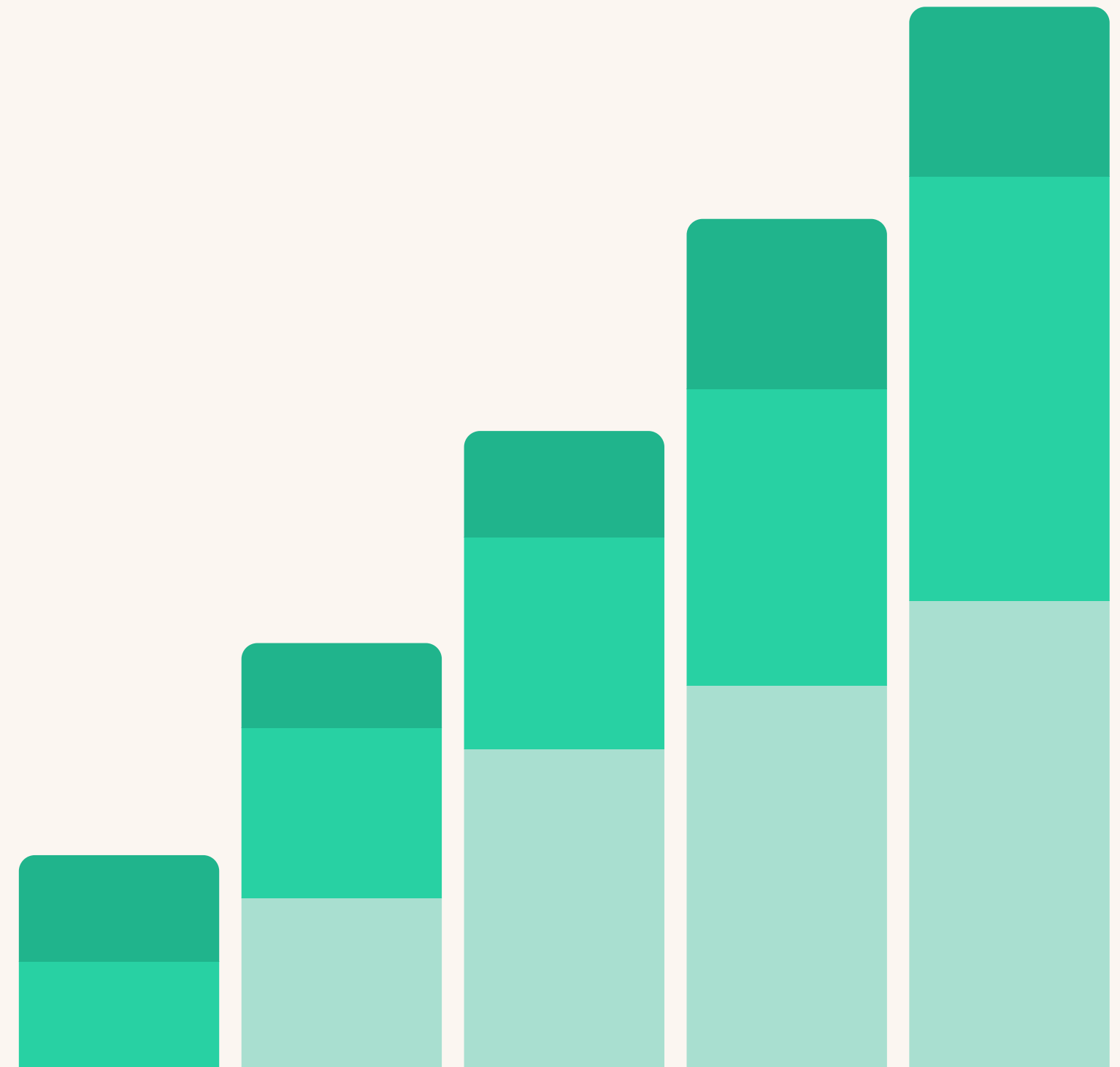


Tujuan

- Mengetahui fitur-fitur mana yang paling berpengaruh terhadap:
 - profit
 - shipping_cost
 - Membandingkan hasil statistik (ANOVA) dan model prediktif
 - Memilih model terbaik dan mengevaluasi performanya
- 

Metodologi

1. Pembuatan model untuk masing-masing variabel target (profit & shipping_cost)
2. Jalankan ANOVA
3. Urutkan fitur/variabel independen berdasarkan signifikansi
4. Buat dataset baru yang berisi variabel target dan variabel independen
5. Pembuatan data train dan test berdasarkan dataset baru
6. Latih model prediktif (Linear, Ridge, RF, XGB, GB)
7. Pilih model terbaik (berdasarkan R^2 /RMSE)
8. Lihat feature importance dari model terbaik
9. Bandingkan hasil dengan ANOVA



ANOVA: shipping_cost

	Feature	F_value	p_value
6	category	2.427126e+03	0.000000e+00
8	order_priority	7.282672e+02	0.000000e+00
7	sub_category	6.717117e+02	0.000000e+00
0	ship_mode	3.939253e+02	5.324038e-253
4	market	9.795918e+01	5.307156e-123
5	region	4.503154e+01	2.172867e-107
3	country	7.698221e+00	1.582053e-148
2	state	1.980443e+00	1.350908e-70
9	year	8.615089e-01	4.602157e-01
1	segment	1.797872e-01	8.354485e-01

- 1.category
- 2.order_priority
- 3.sub_category
- 4.ship_mode
- 5.market
- 6.region
- 7.country
- 8.state



ANOVA :

profit

	Feature	F_value	p_value
6	category	3.042301e+02	4.489104e-132
7	sub_category	1.038995e+02	0.000000e+00
5	region	3.428132e+01	3.218862e-80
4	market	2.824399e+01	6.676329e-34
3	country	2.224210e+01	0.000000e+00
2	state	4.707880e+00	0.000000e+00
8	order_priority	1.569623e+00	1.944132e-01
9	year	2.795162e-01	8.402259e-01
1	segment	2.220602e-01	8.008679e-01
0	ship_mode	1.053396e-01	9.569877e-01

- 1.category
- 2.sub_category
- 3.region
- 4.market
- 5.country
- 6.state



Comparison

Model:
shipping_cost

$y = \text{shipping_cost}$

$x = \text{category, order_priority, sub_category, ship_mode, market, region, country, state + sales, discount, profit, quantity}$

Model:
profit

$y = \text{profit}$

$x = \text{category, sub_category, region, market, country, state + sales, discount, quantity}$

Comparison

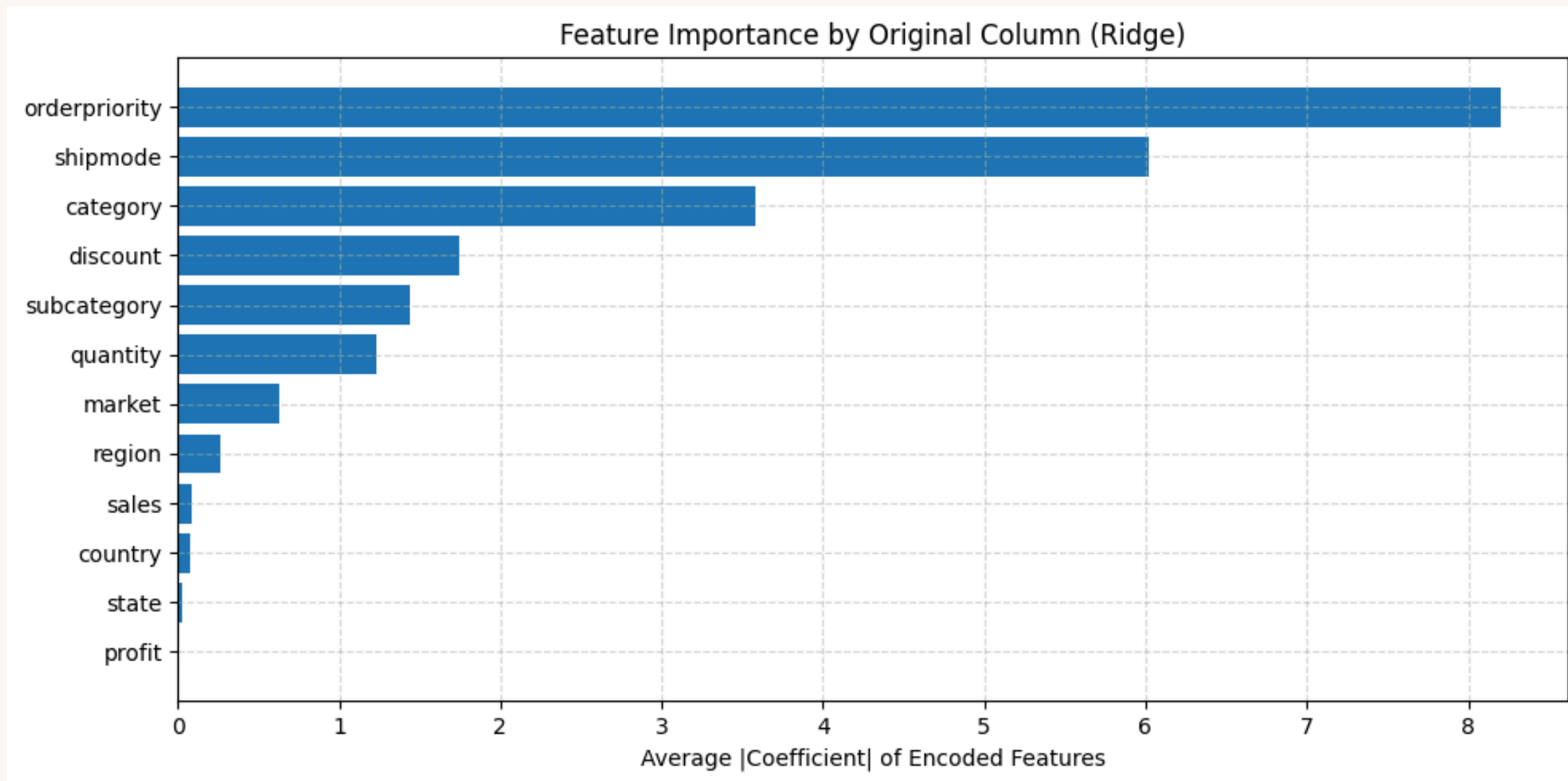
Model: shipping_cost

	Model	MAE	MSE	RMSE	R ² Score
1	Ridge Regression	16.211734	1619.026069	40.237123	0.522528
4	XGBoost	17.958505	1855.159890	43.071567	0.452890
0	Linear Regression	16.365536	1888.459077	43.456404	0.443069
2	Random Forest	15.782310	2096.166315	45.783909	0.381814
3	Gradient Boosting	20.232866	2304.823356	48.008576	0.320278

Model: profit

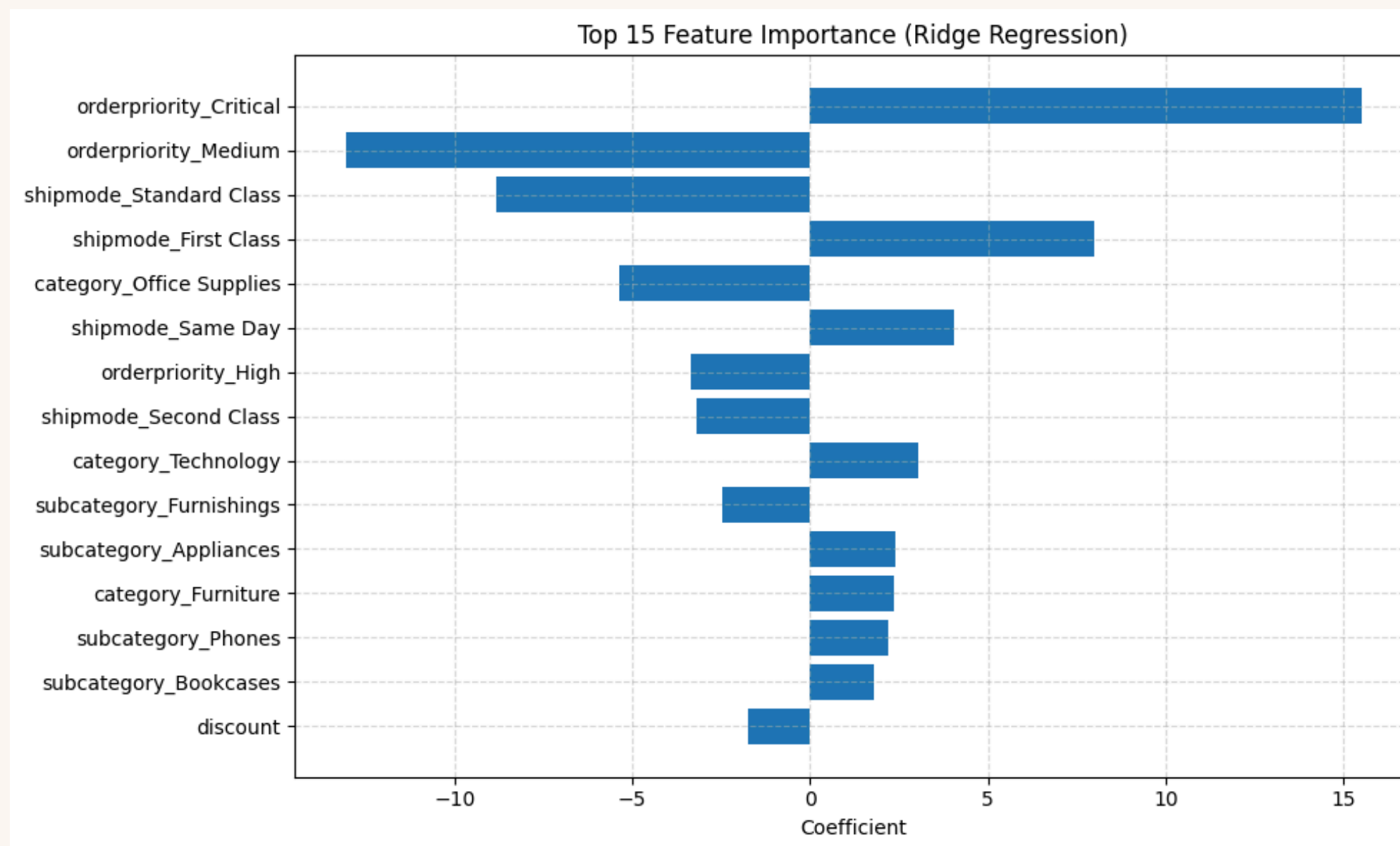
	Model	MAE	MSE	RMSE	R ² Score
2	Random Forest	44.722831	23179.992289	152.249769	0.293645
4	XGBoost	56.130653	24051.809718	155.086459	0.267078
1	Ridge Regression	59.855705	25828.803647	160.713421	0.212928
0	Linear Regression	58.792837	26921.591742	164.078005	0.179628
3	Gradient Boosting	60.845409	28121.084895	167.693425	0.143076

Feature Importance: shipping_cost



- 1.order_priority
- 2.ship_mode
- 3.category
- 4.discount
- 5.sub_category
- 6.quantity
- 7.market
- 8.region
- 9.sales
- 10.country
- 11.state
- 12.profit

Feature Importance: shipping_cost



Comparison: shipping_cost

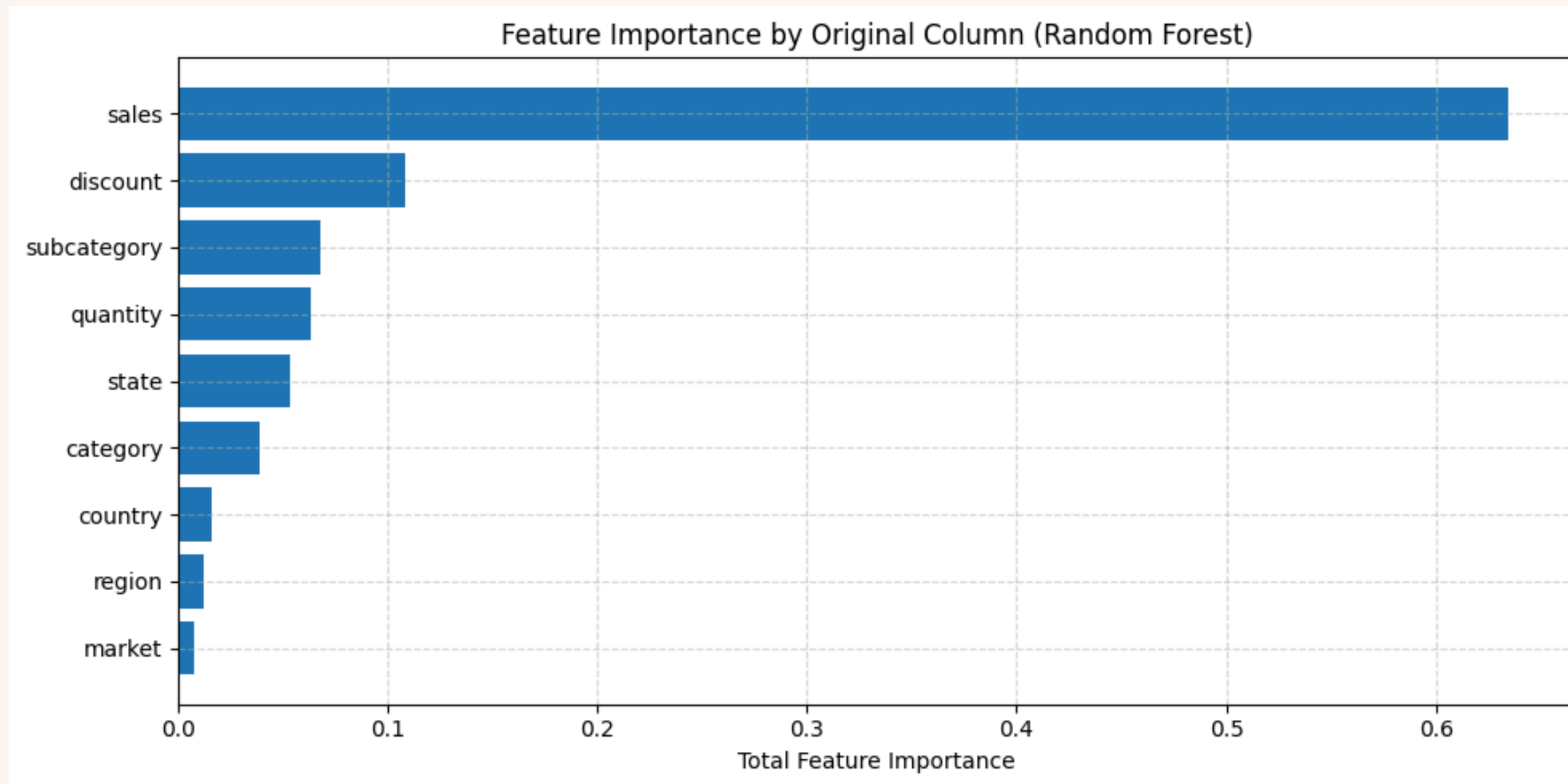
ANOVA

- 1.category
- 2.order_priority
- 3.sub_category
- 4.ship_mode
- 5.market
- 6.region
- 7.country
- 8.state

Ridge Regression

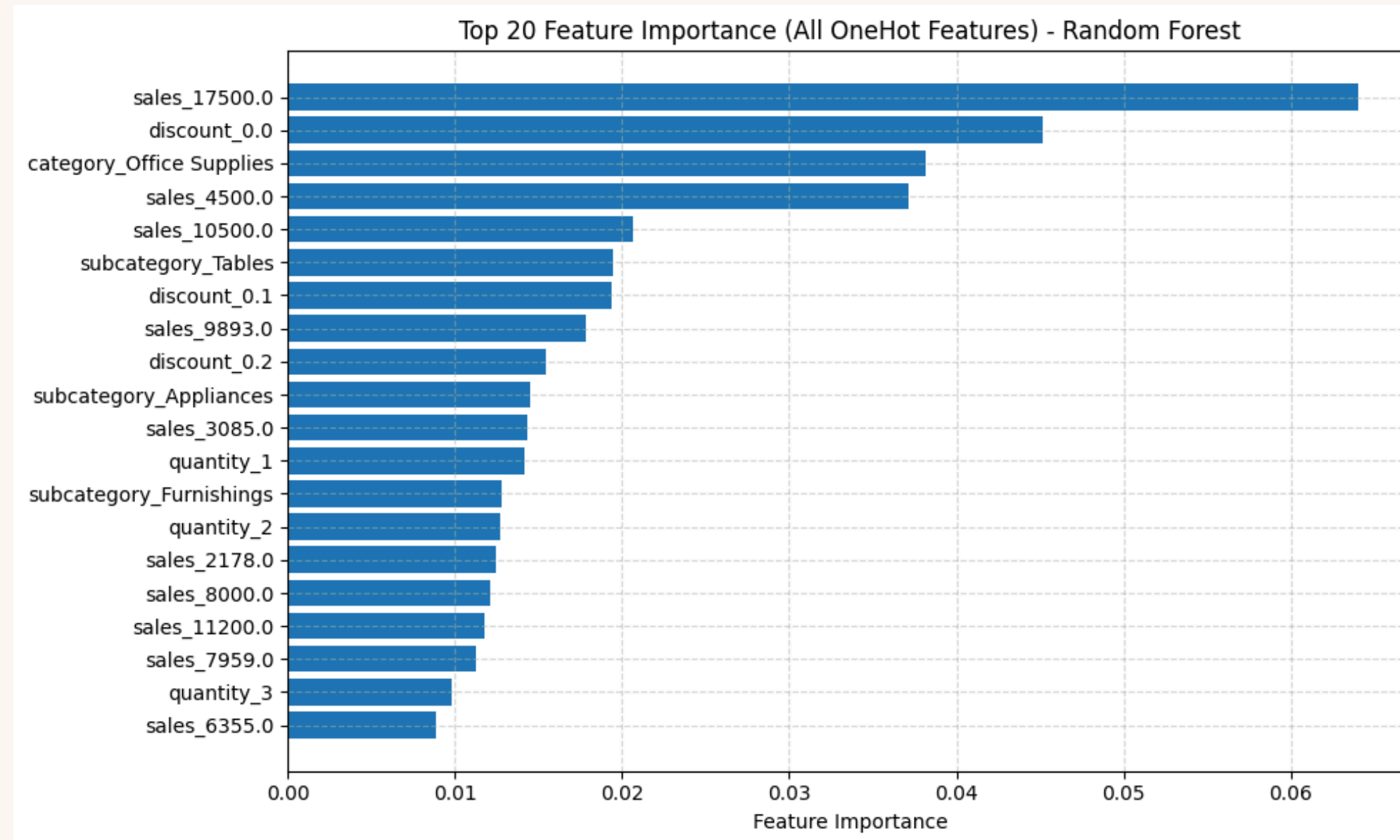
- 1.order_priority
- 2.ship_mode
- 3.category
- 4.discount
- 5.sub_category
- 6.quantity
- 7.market
- 8.region
- 9.sales
- 10.country
- 11.state
- 12.profit

Feature Importance: profit

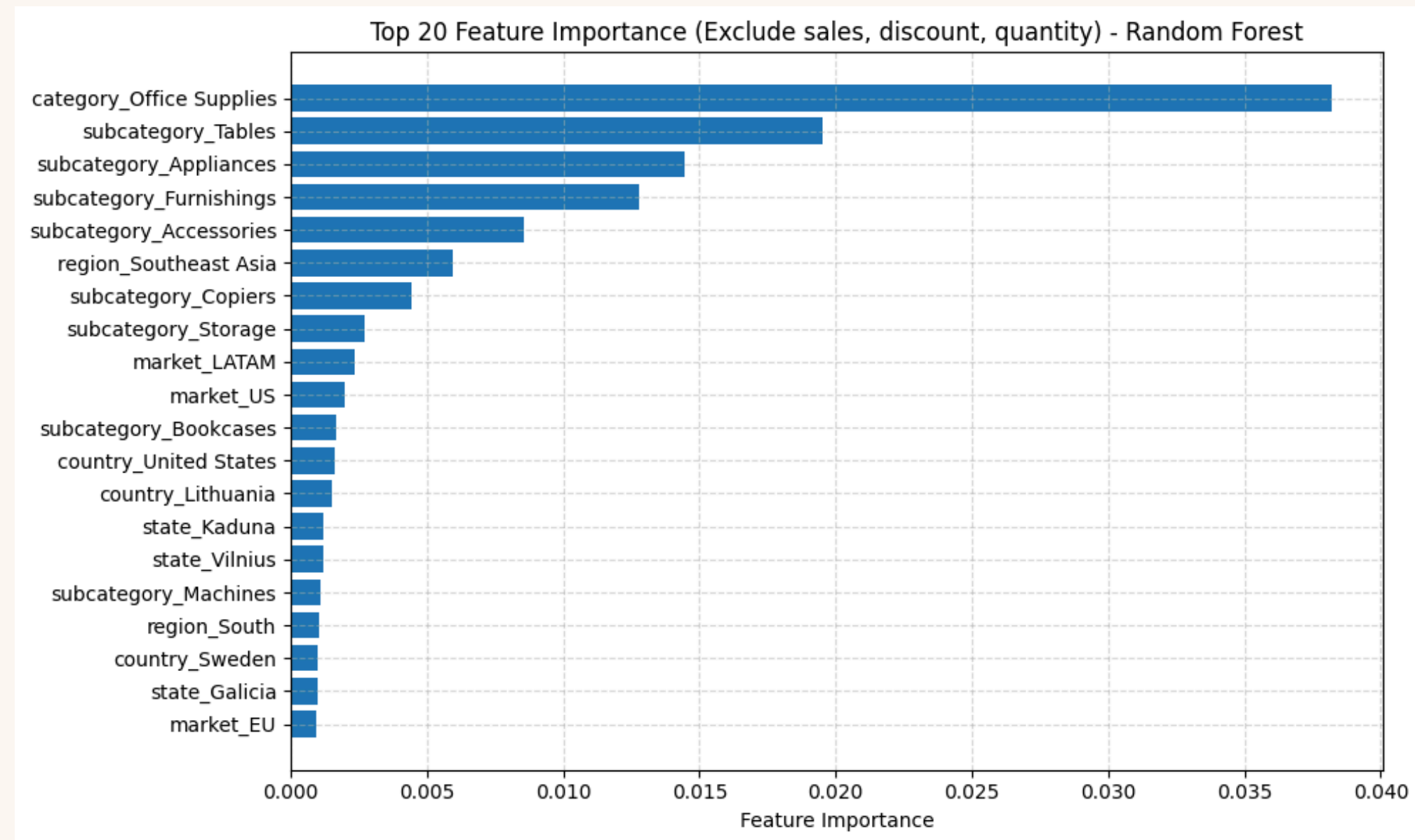


1. sales
2. discount
3. sub_category
4. quantity
5. state
6. category
7. country
8. region
9. market

Feature Importance: profit



Feature Importance: profit



Comparison: profit

ANOVA

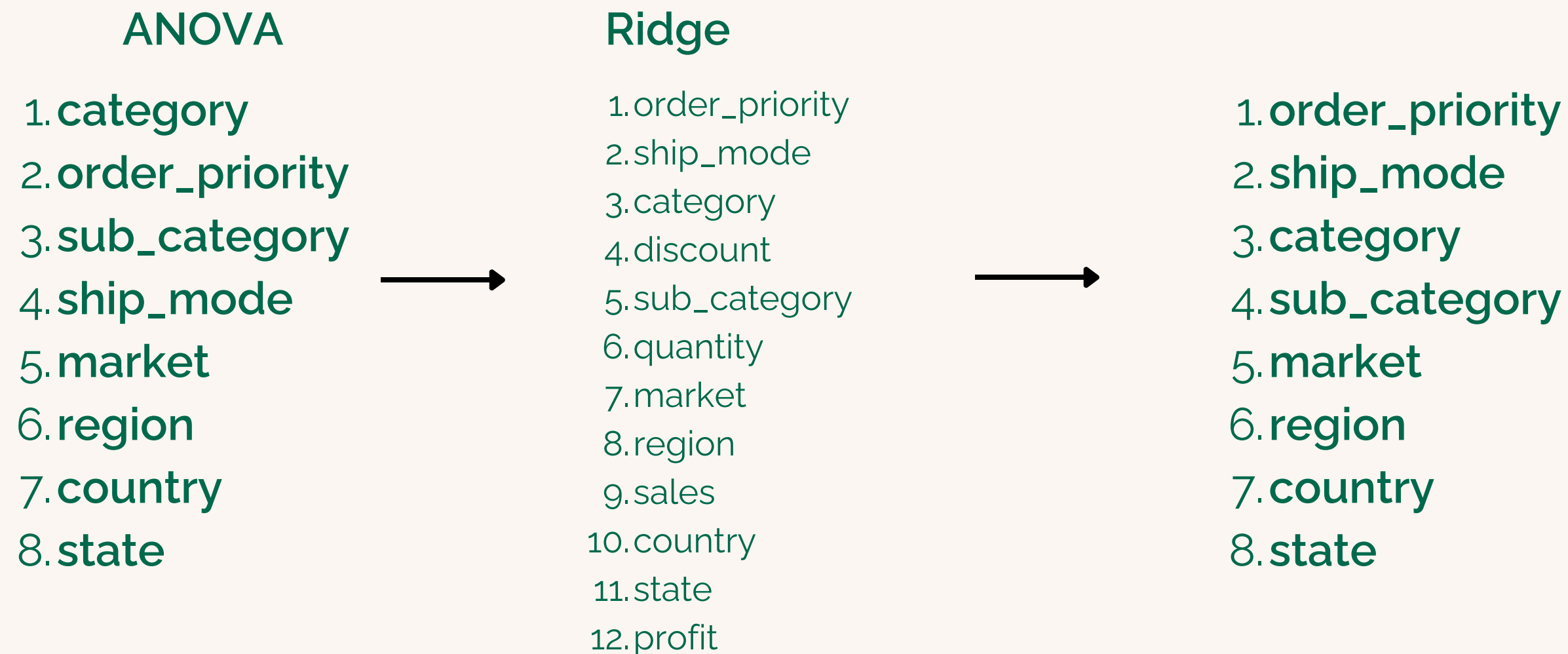
1. category
2. sub_category
3. region
4. market
5. country
6. state

Random Forest

1. sales
2. discount
3. sub_category
4. quantity
5. state
6. category
7. country
8. region
9. market

Kesimpulan (shipping_cost)

1. Fitur-Fitur yang berpengaruh
 - a. shipping_cost: order_priority (Critical +, Medium -), ship_mode (First Class +, Second Class -), category (Office Supplies +, Technology -), discount(-), quantity(+)
2. Perbandingan Hasil ANOVA dan model Prediktif (shipping_cost)
3. Model terbaik: Ridge Regression



Kesimpulan (shipping_cost)

- **order_priority dan ship_mode adalah Penentu Utama Biaya:** Mengirim pesanan "Critical" atau menggunakan "First Class" secara signifikan meningkatkan biaya. Sebaliknya, pesanan "Medium" dan pengiriman "Standard Class" adalah opsi yang jauh lebih murah.
- Untuk mengurangi biaya pengiriman, perusahaan harus mengurangi penggunaan **shipmode_First Class & order_priority_Critical**.
- **Dampak Kategori Produk pada Biaya:** Kategori produk seperti "Office Supplies" dan "Technology" cenderung memiliki biaya pengiriman yang lebih tinggi. Apakah ada cara untuk mengoptimalkan kemasan, rute pengiriman, atau negosiasi dengan penyedia logistik untuk kategori ini?
- **Efek Diskon pada Biaya Pengiriman (Interpretasi Hati-hati):** Koefisien negatif untuk discount menunjukkan bahwa barang yang didiskon cenderung memiliki biaya pengiriman yang lebih rendah.

Kesimpulan (profit)

1. Fitur-Fitur yang berpengaruh
 - a. profit: category (Office Supplies), Region (SEA), sub_category (Tables, Appliances, Furnishings)
2. Perbandingan Hasil ANOVA dan model Prediktif (profit)
3. Model terbaik: Random Forest



Kesimpulan (profit)

- **Fokus Utama pada Harga dan Volume:** Profit sangat didorong oleh seberapa banyak produk terjual (sales), pada tingkat diskon berapa, dan berapa kuantitasnya.
- **Perhatikan Titik Harga dan Diskon Spesifik:** Ada titik-titik penjualan tertentu (\$17,500) dan kebijakan diskon (tanpa diskon) yang sangat menguntungkan.
- **category_Office Supplies, subcategory_Tables, subcategory_Appliances, subcategory_Furnishings, dan subcategory_Accessories** adalah fitur kategorikal teratas. Fokus pada promosi atau peningkatan efisiensi untuk kategori/subkategori yang paling menguntungkan.
- Beberapa region (misalnya **region_Southeast Asia**), market (misalnya **market_LATAM**, **market_US**), country (misalnya **country_United States**, **country_Lithuania**), dan state (misalnya **state_Kaduna**, **state_Vilnius**) juga muncul.

Presented by Michael Abellard Hendra

Thank you very much!

