



Graph-Based Customer Segmentation with GraphSAGE on a Customer–Vehicle Bipartite Network

Abdullah SEZDİ^{1*}, Metin BİLGİN²

¹ arabam.com, <https://orcid.org/0009-0000-5639-7535>, abdullah.sezdi@arabam.com:

² Bursa Uludağ Üniversitesi, Bilgisayar Mühendisliği Bölümü,

Orcid ID: <https://orcid.org/0000-0002-4216-0542>, metinbilgin@uludag.edu.tr:

*Corresponding author, Abdullah.sezdi@arabam.com; 0 541 264 9824

Received: 22 May 2025

Revised: 21 October 2025

2nd Revised: 08 November 2025

Accepted: 18 November 2025

Published: 31 December 2025

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Reference: Sezdi, A., & Bilgin, M. (2025). Customer–vehicle bipartite graph constructed via purchase relationships: GraphSAGE-based edge prediction and embedding-driven customer segmentation. *Orclever Proceedings of Research and Development*, 7(1), 16–29.

Abstract

This study models customer–vehicle interactions in an online used-car platform as a bipartite structure, constructing a graph with customer (U) and vehicle (V) nodes. Relations between the two node sets are defined only by edges representing realized purchase events ($e=(u,v,t)$), thereby focusing on a signal with high business value and relatively low noise. On this graph, inductive node representations (embeddings) are learned with GraphSAGE. During training, link prediction is used solely as a self-supervised proxy task; optimization employs an MLP-based scorer with Binary Cross-Entropy (BCE) loss. Early stopping is triggered when the BCE on a temporally held-out validation set stops improving; together with temporal negative sampling, this prevents leakage of future information.

The objective is to obtain high-quality customer/vehicle embeddings. The learned representations are then used to construct embedding-based customer segments via K-Means. Segmentation quality is evaluated using the Silhouette and Calinski–Harabasz scores. The results show that GraphSAGE embeddings learned on the purchase-induced



bipartite graph provide a practical and scalable foundation for recommendation/targeting and customer understanding tasks.

Keywords: *bipartite graph, purchas, GraphSAGE, link prediction, customer segmentation, K-Means, automotive analytics.*



Konferans Bildirisi

Satın Alma İlişkisiyle Kurulan Müşteri–Araç Bipartit Grafı GraphSAGE Tabanlı Kenar Tahmini ve Gömme Odaklı Müşteri Segmentasyonu

Özet

Bu çalışma, çevrim içi ikinci el otomotiv platformunda müşteri–araç etkileşimini iki parçalı (bipartit) bir yapı olarak ele alır ve müşteri (User-U) ile araç (Vehicle-V) düğümlerinden oluşan bir graph kurar. İki düğüm kümesi arasındaki ilişkiler, yalnızca gerçekleşmiş satın alma olaylarını temsil eden kenarlar ($e = (u, v, t)$) ile tanımlanır; böylece yapı doğrudan yüksek iş değeri taşıyan ve görece düşük gürültülü bir sinyale odaklanır. Bu graph üzerinde GraphSAGE ile induktif düğüm temsilleri (gömmeler) öğrenilir. Öğrenme sürecinde bağlantı tahmini yalnızca özdenetimli bir ön-görev (proxy) olarak kullanılır; MLP tabanlı bir skorlayıcı ve İkili Çapraz Entropi (BCE) kaybı ile optimize edilir. Erken durdurma, zamansal olarak ayrılmış doğrulama kümesi üzerindeki BCE iyileşmeyi bıraktığında uygulanır; böylece gelecek bilgisinin sızması, zamansal negatif örnekleme ile birlikte engellenir. Nihai amaç yüksek kaliteli müşteri/araç gömmeleri elde etmektir: elde edilen temsiller K-Means ile müşterilerin gömme-tabanlı segmentasyonunda kullanılır. Segmentasyon kalitesi Silhouette ve Calinski–Harabasz skorlarıyla değerlendirilmiş; sonuçlar, satın alma ilişkisiyle kurulan bipartit grafik üzerinde öğrenilen GraphSAGE temsillerinin öneri/hedefleme ve müşteri anlayışı görevleri için pratik ve ölçeklenebilir bir temel sunduğunu göstermiştir.

Anahtar: bipartit graf, satın alma, GraphSAGE, link prediction, müşteri segmentasyonu, K-Means.



1. Giriş

Çevrim içi ikinci el otomotiv pazarında müşteri ile araç arasındaki satın alma olayları seyrek ve zamana duyarlıdır. Tablo odaklı yaklaşımlar komşuluk ve yapı bilgisini sınırlı yansıttığı için belirli bir müşteri–araç çifti için satın alma olasılığını kestirmek ve müşterileri davranışsal olarak ayırtırmak zorlaşır. Bu çalışmada müşteri (U) ve araç (V) düğümlerinden oluşan bipartit bir graph kurulmakta ve kenarları yalnızca gerçekleşmiş satın alma olayları olarak şekilde tanımlamaktayız. Çalışmanın zorlukları;

- **Seyreklik:** Satın alma nadirdir; negatif örnek seçimi kritiktir.
- **Zaman etkisi:** Gelecek bilgisinin sızması (leakage) metrikleri yapay biçimde şişirebilir.
- **Yüksek kardinalite:** Araç ve kullanıcı alanları geniştir; temsillerin genellemesi gerekir.
- **İşe çevrim:** Temsillerin segment/stratejiye dönüşmesi izlenebilir olmalıdır.

şeklinde sayılabilir.

Yukarıda sözü edilen problemin ve zorlukların çözümü için geliştirilen yaklaşım kısaca şöyle özetlenebilir. Öncelikle, Bipartit graph üzerinde GraphSAGE ile indüktif düğüm temsilleri (gömmeler) öğreniliyor. Ardından, öğrenme sinyali olarak bağlantı tahminini (link prediction) self-supervised bir ön görev şeklinde optimize ediliyor; (u, v) satın alma tahmini, yüksek kaliteli müşteri/araç temsilleri öğrenmek için vekil bir kayıp (proxy loss) işlevi görüyor. Elde edilen temsiller, K-Means ile gömme tabanlı müşteri segmentasyonu ve analitik benzerlik/sıralama çalışmaları için kullanılıyor. Eğitim kurgusunda zamana dayalı ayırım (train/validation/test) ve zamansal negatif örnekleme ile sızıntı önleniyor. Çalışmanın katkıları olarak;

- Satın alma kenarıyla tanımlı bipartit şema ve zamana duyarlı değerlendirme,
- Self-supervised GraphSAGE: bağlantı tahminini vekil kayıp olarak kullanıp genellenebilir gömmeler,
- Tek temsille çok görev: segmentasyon ve analitik sıralama/benzerlik,
- Uzatılabilirlik: çoklu etkileşim türleri ve ranking loss ile genişlemeye uygun mimari,



şeklinde sayılabilir.

1.1.İlişkili Çalışmalar

Bu bölümde ilgili alanda yapılmış çalışmalara ait bilgiler sunulmuştur. Klasik öneri ve ağ gömme ile ilgili çalışmalar; Kullanıcı-öğre etkileşimlerinde matris çarpanlarına ayırma ve BPR, etkileşim matrisinden düşük boyutlu temsiller türeterek sıralama performansı sağlamıştır (Rendle, 2009). Rasgele yürüyüş ve bağlamsal birlikte-görünmeye dayalı yöntemler (DeepWalk/Node2Vec) ile öğre/kullanıcı gömmeleri üretilmiştir (Grover & Leskovec, 2016). Bu aile komşuluk yapısını örtük işler; bipartit topolojide indüktif genelleme ve zaman duyarlı değerlendirme sınırlı kalabilir. Çalışmamız, bipartit yapıda doğrudan komşuluk mesajlaşması ve zamansal ayırım ile bu sınırlara çözüm getirmektedir.

GNN tabanlı öneri ve bağlantı tahmini ile ilgili çalışmalar; GraphSAGE, örneklemeli komşuluk toplamalarıyla büyük graflarda indüktif temsil öğrenimi sunmaktadır (Hamilton ve ark., 2017). Bipartit öneri senaryolarında PinSAGE, örnekleme ve ölçeklenebilirlik konularını ele almıştır (Ying ve ark., 2018); kullanıcı-öğre etkileşimi için LightGCN/NGCF hattı, mesajlaşma mimarisini sadeleştirmiştir (Wang ve ark., 2019; He ve ark., 2020).

Bu çalışmalardan farklı olarak, yaptığımız çalışmada yalnızca satın alma kenarını kullanırken, bağlantı tahminini nihai amaç değil proxy sinyal olarak optimize etmekte ve çıktıyı müşteri segmentasyonuna taşımaktayız.

Self-supervised GNN ve segmentasyon ile ilgili çalışmalar; GNN'lerde self-supervised amaçlar (ör. BCE/InfoNCE, veri artırma) temsillerin genellenebilirliğini güçlendirmiştir (Wu ve ark., 2021; You ve ark., 2020). Bu çalışmada ise MLP + BCE ile optimize edilen bağlantı tahmini sinyalini zaman duyarlı ayırımla birleştirir; elde edilen temsilleri K-Means ile gömme tabanlı müşteri segmentasyonunda değerlendirir. Segmentasyon kalitesi Silhouette ve Calinski-Harabasz skorlarıyla raporlanır (Rousseeuw, 1987; Calinski & Harabasz, 1974).

İlişkili çalışmalar ve yapılan çalışmanın karşılaştırılması Tablo 1'de sunulmuştur.



Tablo 1 İlişkili Çalışmalar ile karşılaştırma (bipartif, indüktif, self supervised)

Yöntem	Bipartit	İndüktif	Self-sup.	Kaynak
BFR/MF	✓	-	-	Rendle (2009); Koren, Bell & Volinsky (2009)
Node2Vec	✓	kısmi	-	Grover & Leskovec (2016)
PinSAGE	✓	✓	-	Ying et al. (2018)
LightGCN/NGCF	✓	kısmi	-	Wang et al. (2019); He et al. (2020)
Çalışmamız	✓	✓	✓(LP proxy)	

2. Materyal ve Yöntem

Bu bölümde çalışmada kullanılan yöntemler ile ilgili ayrıntılı bilgiler sunulmuştur.

2.1. Graph Kurulumu ve Kapsam

Öncelikle belirli bir tarih aralığındaki (ör: 01/2020–06/2025 dönemindeki) 114.295 gerçekleşmiş satın alma işlemi temelinde iki parçalı bir graf $G = (U \cup V, E)$ kurulmaktadır. Bu graptta; U , 5.631 benzersiz müşteriye, V ise 4.819 farklı araç modeli düğümünü temsil etmektedir. Kenar kümesi E , yalnızca satın alma ilişkilerini ve zaman damgasını içerir ($e = (u, v, t)$); ve görüntüleme/favori/teklif gibi diğer etkileşimler kapsam dışı tutulmaktadır. **Model düğümü gerekçesi** Tekil araç ID'si yerine *model* düzeyi (i) sıklığı azaltır, (ii) gömme genellemesini artırır, (iii) iş tarafında yorumlanabilirliği korur.

2.2. Kullanılan Özellik Kümeleri

Bu alt bölüm, müşteri (U) ve araç (V) düğümlerinde kullanılan özellikleri mantıksal gruplar halinde özetler. Bu çerçeve (i) projeksiyon katmanlarının girdilerini netleştirir, (ii) model yorumlanabilirliğine katkı sağlar, (iii) veri kalitesini kategori bazında izlemeyi kolaylaştırır. Grupların tam dökümü Tablo 2 ve Tablo 3'te sunulmuştur.



Table 2 Müşteri (U) özellik kategorileri.

Kategori	Kısa açıklama /örnekler	Tür
İşlem hacmi & değer	Toplam işlem,ort işlem tutarı,işlem başına ort teklif adedi	Sayısal
Marka / İhale Çeşitliliği	Marka yoğunluğu (Gini), marka/ihale entropisi	Sayısal
Zamansallık	Son alış üzerinden gün (recency), müşteri yaşı (tenure)	Sayısal
Coğrafya	Farklı şehirden alım(0/1),farklı şehir sayısı, kendi şehirden araç alma oranı	İkili / Sayısal
Üyelik / Plan	Paket Sahipliği (0/1)	İkili
Araç yaş tercih dağılımı	0-1,2-4,5-6,7-10,10+ yaş alım payları (Müşterinin tüm satın alımları içindeki oranlar)	Oran vektörü
Çeyreklere göre alım yoğunluğu	2024 Q1 – 2025 Q2 arası çeyreklerine göre alım sayıları	Sayısal

Table 3 Araç (V) özellik kategorileri.

Kategori	Kısa açıklama	Tür
Teknik/Kategorik	Yakıt türü, Vites Türü, Araç tipi, Marka	Kategorik
Model istatistikleri	Model bazında ort.satış fiyatı (USD),model bazında ort. Teklif sayısı	Sayısal
Yaş	Aracın alındığı zamanki aracın yaşı	Sayısal

2.3. Ön-İşleme ve Özelliklerin Hizalanması

Bu kısımda yapılan ön işlemler ve özelliklerin hizalanması ile ilgili bilgiler sunulmuştur.

- **Sayısal özellikler:** StandardScaler ile standartlaştırılmıştır; ölçekleyiciler yalnızca **eğitim** penceresinde fit edilip doğrulama/testte aynen kullanılmıştır.
- **Kategorik özellikler:** Her alan için eğitilebilir gömme (trainable embedding) kullanılmıştır.
- **Tipler arası hizalama:** Müşteri (U) ve araç (V) sayısal vektörleri, tip-spesifik öğrenilebilir projeksiyonlarla ortak hedef boyuta eşlenmiş ve kategorik alanlar, eğitim sırasında bu temsillere birleştirilmiştir.

2.4. GraphSAGE ile Mesajlaşma

SAGE, örneklemeli komşuluk toplamalarıyla *indüktif* öğrenme sağlamaktadır [3].

Bipartit grafimizde mesajlaşma $U \leftrightarrow V$ yönündedir; toplayıcı=*mean*, $L=2$ olacak şekilde Eşitlik 1'de sunulmuştur.

$$m_u^{(1)} = \frac{1}{|\mathcal{N}(u) \cap V|} \sum_{w \in \mathcal{N}(u) \cap V} h_w^{(1)}, \quad h_u^{(l+1)} = \sigma \left(W_U^{(l)} \left[h_u^{(l)} \parallel m_u^{(l)} \right] \right) \quad (1)$$

v için simetrik; son temsiller $z_u = h_u^{(L)}, \quad z_v = h_v^{(L)}$

2.5. Bağlantı Tahmini

Bu kısımdaki amaç gömme kalitesini dolaylı izlemek için *satın alma* bağıntı tahmin eden bir *ön-görev* (*proxy*) optimize etmektir. Kenar temsili ve skor hesabı için Eşitlik 2-3 sunulmuştur.

$$\varphi(u, v) = [z_u \parallel z_v] \in \mathbb{R}^{2d}, \quad s(u, v) = g_{\theta}(\varphi(u, v)) \quad (2)$$



burada $g_{\phi}(\cdot)$ iki gizli katmanlı bir MLP'dir. Olasılık, lojistik sigmoid ile:

$$\hat{y}(u, v) = \sigma(s(u, v)) = \frac{1}{1 + \exp(-s(u, v))} \quad (3)$$

Dengesiz sınıf için pozitif ağırlıklı ikili çapraz entropi (Sigmoid+BCE) için Eşitlik 4 kullanılır:

$$\bullet \quad L = -\alpha y \log(\hat{y}) - (1 - \alpha)(1 - y) \log(1 - \hat{y}), \quad y \in \{0,1\} \quad (4)$$

α pozitif sınıf ağırlığıdır (pos_weight). Negatif örnekler, her pozitif için **3:1** oranında ve **aynı zaman penceresinden zamansal negatif örnekleme** ile üretilir; geçmişte pozitifleşmiş (u,v) çiftleri hariç tutularak gelecek bilgisi sızıntısı ve yanlış-negatif riski azaltılır.

2.6. Değerlendirme Protokolü

Bu bölümde, Temporal split kullanılarak değerlendirme yapılmıştır. Eğitim için 01/2020–12/2024 aralıkları, Doğrulama için 01/2025–03/2025 aralıkları ve Test için 04/2025–06/2025 aralıkları kullanılmıştır. Her pencerede gömmeler pencere başlangıcından önceki snapshot'tan üretilir; değerlendirme kenarları ilgili pencere içinde hesaplanır.

2.1.Uygulama

Bu çalışmada iki katmanlı GraphSAGE mimarisi ($L=2$, toplayıcı=mean) kullanılarak $d=128$ boyutlu gömmeler öğrenildi; üstüne 256 gizli birimli, ReLU+Dropout içeren bir MLP ile bağlantı olasılığı modellendi. Negatif örnekleme, her pozitif kenar için aynı zaman penceresinden 3 negatif üretilecek şekilde uygulandı; sınıf dengesizliği $\text{pos_weight} = \alpha$ ile dengelendi. Eğitim süreci, zamansal olarak ayrılmış doğrulama kümesi üzerindeki BCE kaybı izlenerek ve kayıp iyileşmeyi bıraktığında erken durdurma uygulanarak yönetildi. Nihai amaç gömme kalitesinin segmentasyonda karşılık bulması olduğundan, müşteri gömmeleri K-Means ile $K \in [2,19]$ aralığında kümelendi; Silhouette

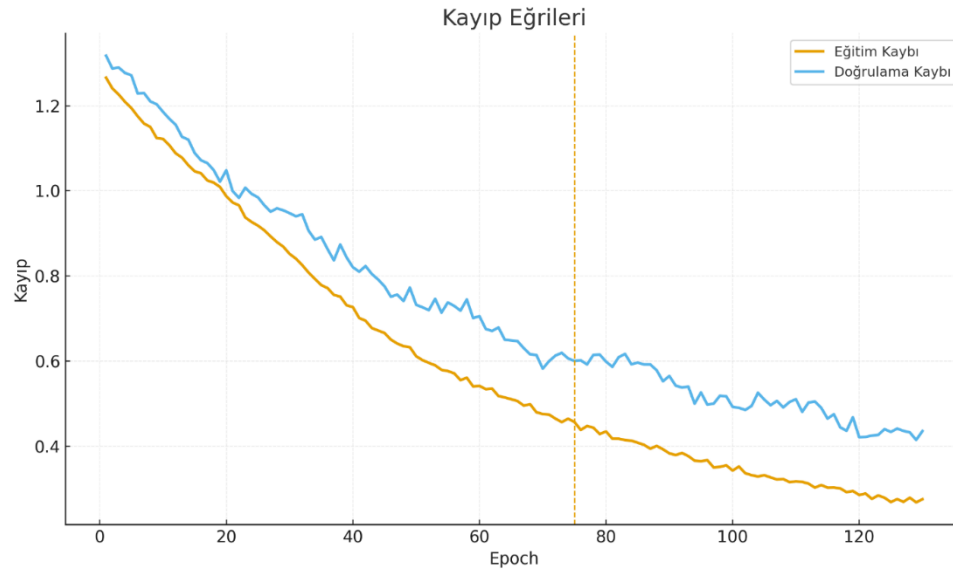
ve Calinski–Harabasz ölçütlerinde görülen yerel plato ile iş birimi yorumlanabilirliği birlikte değerlendirilerek nihai K seçildi

3. Sonuçlar

Bu bölümde (i) proxy bağlantı tahmini performansı ve (ii) gömme tabanlı müşteri segmentasyonunun küme kalitesine ait bilgiler sunulmuştur. Link prediction metrikleri, gömme öğreniminin yardımcı göstergeleridir; asıl amaç, segmentasyon için ayırt edici müşteri temsilleri elde etmektir.

3.1. Bağlantı tahmini

Bağlantı tahmini yalnızca kendinden gözetimli bir “proxy” olarak kullanılmıştır. Eğitim boyunca model gidişatı için tek seçim ölçütü, **zamansal olarak ayrılmış doğrulama kümesi üzerindeki BCE kaybıdır**. Erken durdurma, doğrulama kaybı azalmayı bıraktığında uygulanmıştır (Şekil 1). Kayıp eğrileri ~75. epok civarında plato yapmış, aşırı uyum sinyaliyle eğitim sonlandırılmıştır. Nihai hedef, müşteri gömmelerinin **segmentasyon** görevinde ayrıştırıcı ve yorumlanabilir kümeler üretmesidir; bu nedenle ayrıntılı link-prediction karşılaştırmaları sunmuyoruz. Kalite değerlendirmesi bir sonraki bölümde **Silhouette** ve **Calinski–Harabasz** ölçütleri ile ve örnek küme profilleriyle verilmiştir.



Şekil 1 Bağlantı tahmini proxy görevi için eğitim ve doğrulama BCE kayıp eğrileri

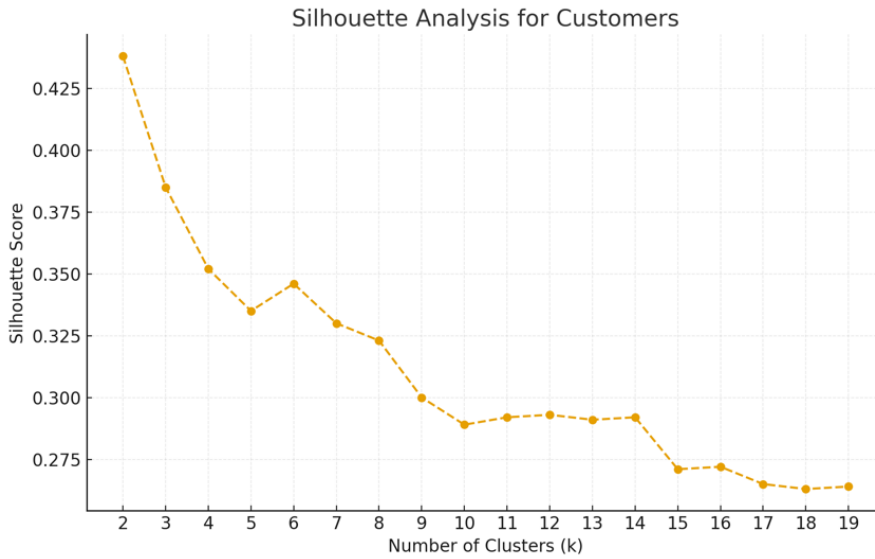


3.2. Gömme-Tabanlı Müşteri Segmentasyonu

Öğrenilen müşteri gömmeleri K-Means ile $K \in [2,19]$ aralığında taranmış; Silhouette ve Calinski-Harabasz skorlarındaki yerel plato ve iş birimi geri bildirimi birlikte değerlendirilerek $K=6$ seçilmiştir. Tablo 4'te ve Şekil 2'de elde edilen sonuçlar sunulmuştur.

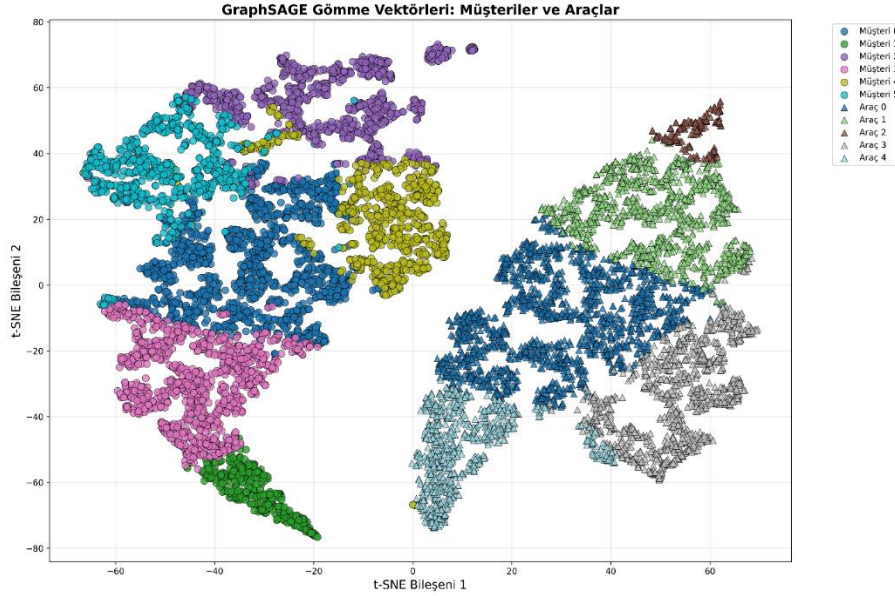
Table 4 Segmentasyon kalite metrikleri (müşteri gömmeleri).

K	Silhouette	Calinski-Harabasz
3	0.385	3000
4	0.353	2700
5	0.336	2450
6	0.346	2550
7	0.331	2350



Şekil2 Silhouette eğrisi ($k=2-19$); $K \approx 6$ civarında yerel plato.

$K=2$ maksimum Silhouette'i verse de küme sayısı iş kullanım senaryoları için fazla kaba kaldığı gözlemlenmiştir.. $K=6$, ayrışma (Silhouette) ile kompaktlık (CH) arasında dengeli bir nokta sunarken iş birimlerinin talep ettiği granüleriteyi sağlar.



Şekil 3 Müşteri ve Araç Gömmelerinin t-SNE İle 2D Projeksiyonu

Şekil 2, K arttıkça ayrışmanın nasıl azalarak plato yaptığını gösterir. Şekil 3, müşteri ve araç temsillerinin 2B izdüşümünde kümelerin birbirinden ayrılabilirdiğini (tamamen ayırma beklenmeden) görsel olarak destekler.

Çalışma sonunda elde edilen çıkarımlar;

- Proxy sinyalin yeterliliği: LP metriklerindeki istikrarlı trend, gömme kalitesinin segmentasyon çıktılarına yansıdığını göstermektedir.
- Küme kalitesi: $K=6$ seçimi, Silhouette/CH ile uyumlu ve iş birimi açısından yorumlanabilir bir granülerite sunmaktadır.
- Pratik aktarım: Öğrenilen temsiller, segment-temelli hedefleme ve analitik benzerlik/sıralama görevleri için yeniden kullanılabilir bir temel sağlamaktadır.

4. Tartışma

Çalışma sonuçlarının değerlendirildiğinde;

Güçlü yönler: Gürültüsü düşük satın alma sinyali, (ii) indüktif gömme ile soğuk başlangıç toleransı, (iii) tek temsil ile çok-görev (öneri/benzerlik/segment) şeklinde,



Sınırlılıklar: (i) Görüntüleme/favori/teklif gibi öncül etkileşimler dışında; (ii) zaman etkisi *time-decay* ile açıkça modellenmedi; (iii) proxy hedef ile iş hedefleri arasında dolaylı köprü şeklinde söylenebilir.

Gelecek çalışmalarına ait ise çok-ilişkili/heterojen GNN (view/favori/teklif+satın alma), öneri odaklı *ranking* kayıpları (BPR/listwise), kontrastif self-supervision (InfoNCE), içerik/metadata zenginleştirmesi ve canlı A/B şeklinde sunulabilir.



Referanslar

- [1] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009.
- [2] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*, 2016.
- [3] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017.
- [4] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *KDD*, 2018.
- [5] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *SIGIR*, 2019.
- [6] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *SIGIR*, 2020.
- [7] Z. Wu *et al.*, "A comprehensive survey on graph contrastive learning," *IEEE TPAMI*, 2021, arXiv:2106.05264.
- [8] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *NeurIPS*, 2020.
- [9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [10] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27