Müşteri Segmentasyonu Projesi

İstanbul Arel Üniversitesi – Veri Madenciliği Dersi Final Projesi

Oğuzhan Cem Yücel (oguzhancemyucel21@istanbularel.edu.tr)
Can Çebi (cancebi21@istanbularel.edu.tr)
Danışman/Hoca: Sibel Birtane Akar

28 Aralık 2024

1 Giriş

Bu projede, **Kaggle** ortamında yayınlanan "Customer Segmentation" veri seti kullanılmıştır. Projenin temel amacı, bir müşteri kitlesini ortak özelliklerine göre gruplara (Segmentlere) ayırmaktır. Veri madenciliğinde "müşteri segmentasyonu" (customer segmentation), pazarlama stratejileri ve müşteri yönetimi için önemli bir yaklaşımdır. Segmentasyon, pazarlama kampanyalarının özelleştirilmesine, müşteri memnuniyetinin artırılmasına ve kaynakların verimli kullanılmasına katkı sağlar.

Rapor, *İstanbul Arel Üniversitesi*, *Veri Madenciliği* dersi kapsamında hazırlanmıştır. Burada, hem **EDA (Keşifsel Veri Analizi)** hem de **modelleme** adımları sergilenmekte; elde edilen sonuçlar görsellerle birlikte sunulmaktadır.

2 Veri Seti ve Özellikleri

Kullanılan veri seti, Kaggle'dan Customer Segmentation isimli kaynaktan alınmıştır.

- **Train.csv** dosyası: 8068 satır, 11 sütun
- **Test.csv** dosyası: 2627 satır, 10 sütun
- **sample_submission.csv**: Örnek Kaggle gönderim formatı (ID, Segmentation)

Bazı önemli değişkenler şöyledir:

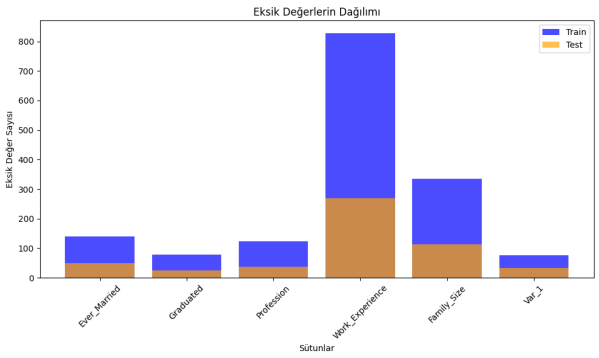
- **ID**: Müşteri ID'si (benzersiz)
- **Gender**: Kadın / Erkek
- **Ever_Married**: Evli olup olmadığı (Yes/No)
- **Age**: Yaş (sayısal)
- **Graduated**: Üniversite mezunu mu (Yes/No)
- **Profession**: Meslek (Artist, Doctor, Engineer, vb.)
- **Work_Experience**: Kaç yıllık iş tecrübesine sahip (sayısal)
- **Spending_Score**: Harcama skoru (Low, Average, High)
- **Family_Size**: Ailedeki kişi sayısı (sayısal)
- **Var_1**: Ek bir kategorik sütun (Cat_1, Cat_2, vb.)
- **Segmentation**: Hedef değişken (A, B, C, D) (Sadece Train.csv'de)

Amaç, Train verisiyle bir model kurup Test verisindeki her müşteri için en olası segmenti (A, B, C, D) tahmin etmektir.

3 Eksik Değer Analizi ve Veri Hazırlama

3.1 Eksik Değerler

Train ve Test veri setlerinde *Ever_Married*, *Graduated*, *Profession*, *Work_Experience*, *Family_Size*, *Var_1* gibi sütunlarda eksik değerler gözlenmiştir. Eksik değerlerin veri setindeki dağılımını aşağıdaki Şekil 1 göstermektedir.



Şekil 1: Eksik Değerlerin Dağılımı

Eksik değerleri doldurma stratejisi aşağıdaki gibidir:

- Numerik sütunlarda eksik değerler, sütunun ortalama (*mean*) değeri ile doldurulmuştur.
- Kategorik sütunlarda eksik değerler, sütunun mod (*mode*, en sık tekrar eden değer) değeri ile doldurulmuştur.
- Eksik değerlerin doldurulması sonrası veri setinde artık eksik değer bulunmamaktadır.

3.2 Veri Dönüşümleri

Veri seti üzerinde yapılan dönüşümler aşağıdaki gibidir:

- Kategorik sütunlar (*Gender*, *Ever_Married*, *Graduated*, *Profession*, *Spending_Score*, *Var_1*), makine öğrenmesi yöntemleri için **Label Encoding** kullanılarak sayısal değerlere dönüştürülmüştür. Örneğin, *Gender* sütununda *Male* değeri 0, *Female* değeri 1 olarak kodlanmıştır.
- Sayısal değişkenler (*Age*, *Work_Experience*, *Family_Size*), veri setindeki ölçek farklılıklarını gidermek ve algoritmaların daha verimli çalışmasını sağlamak amacıyla **StandardScaler** ile ölçeklendirilmiştir.
- Veri dönüşümleri öncesi ve sonrası değerlerin karşılaştırması Tablo 1'da sunulmaktadır.

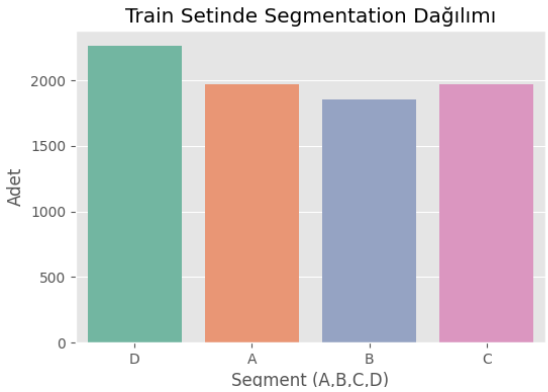
Tablo 1: Veri Dönüşümleri Öncesi ve Sonrası Örnek

Değişken	Dönüşüm Öncesi	Dönüşüm Sonrası
Age	45	0.093
Work_Experience	10	2.165
Family_Size	4	0.756

4 EDA (Keşifsel Veri Analizi) ve Bulgular

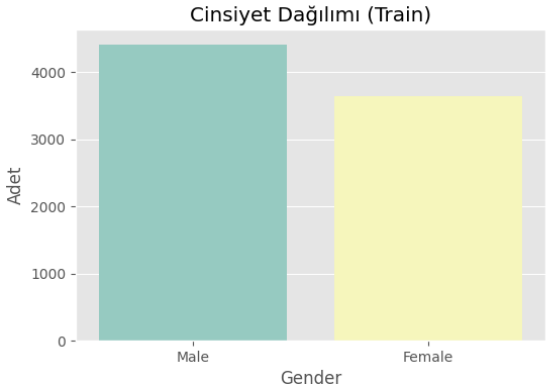
Projemizde **6 adet** grafik oluşturulmuştur:

- **Train Setinde Segmentation Dağılımı**: A, B, C, D segmentlerinin Train verisindeki dağılımını gösterir. Segmentler arasında çok büyük dengesizlik olmadığı görülür.



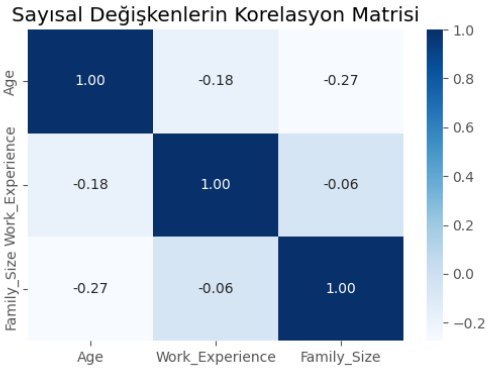
Şekil 2: Train Setinde Segmentation Dağılımı

- **Cinsiyet Dağılımı (Train)**: Erkek ve kadın müşteri sayısı birbirine yakındır.



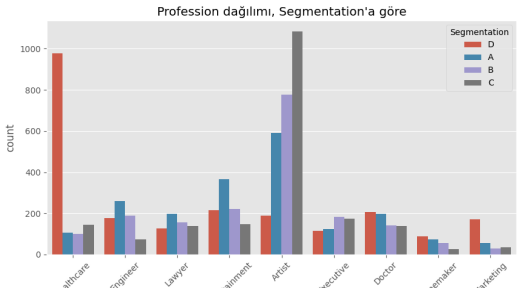
Şekil 3: Cinsiyet Dağılımı (Train)

- **Sayısal Değişkenlerin Korelasyon Matrisi**: Age, Work_Experience, Family_Size arasındaki korelasyon değerlerinin çok yüksek olmadığı gözlemlenir.



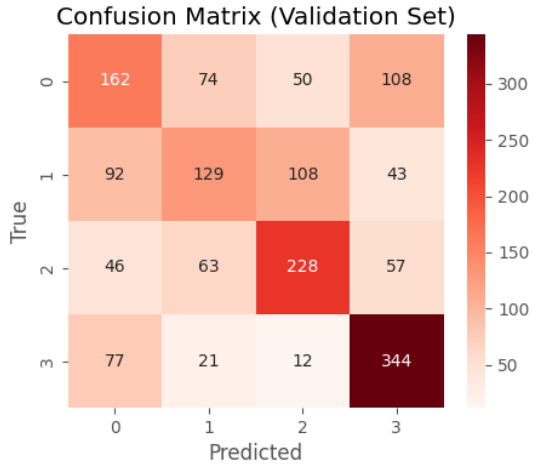
Şekil 4: Sayısal Değişkenlerin Korelasyon Matrisi

- **Profession dağılımı Segmentation'a göre**: Bazı meslek gruplarında, belli segmentlerin (A/B/C/D) yoğunlaştığı görülebilir.



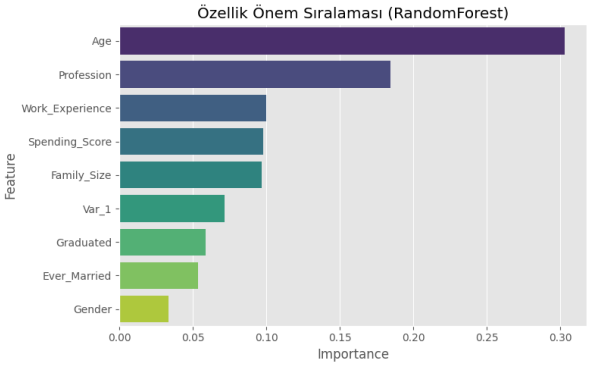
Şekil 5: Profession dağılımı Segmentation'a göre

- **Confusion Matrix (Validation Set)**: Confusion Matrix, modelin Segmentasyon "3" sınıfında yüksek doğruluk sağladığını, ancak diğer sınıflar arasında karışıklık olduğunu göstermektedir.



Şekil 6: Confusion Matrix (Validation Set)

- **Özellik Önem Sıralaması (RandomForest)**: "Age" ve "Profession" özellikleri, segmentasyon tahminlerinde en belirleyici faktörlerdir.



Şekil 7: Özellik Önem Sıralaması (RandomForest)

5 Model Kurulumu, Seçimi ve Eğitimi

5.1 Model Seçimi

RandomForestClassifier tercih edilmiştir. Hem kategorik hem de numerik verilerde iyi sonuç verebilmesi ve *ensemble* yaklaşımına sahip olması büyük avantaj sağlamaktadır. **GridSearchCV** ile:

- n_estimators: [100, 200]
- max_depth: [None, 10, 20]

taranmış, en iyi parametreler {`max_depth: 10, n_estimators: 200`} olarak bulunmuştur.

5.2 Eğitim ve Değerlendirme

Train verisi, `train_test_split` yöntemiyle %80 eğitim, %20 doğrulama şeklinde ikiye ayrılmıştır. Değerlendirme metrikleri: *accuracy, precision, recall, f1-score*.

Validation Sonuçları (RandomForest, en iyi model):

	precision	recall	f1-score	support
0 (A)	0.43	0.41	0.42	394
1 (B)	0.45	0.35	0.39	372
2 (C)	0.57	0.58	0.58	394
3 (D)	0.62	0.76	0.68	454
accuracy			0.53	1614
macro avg	0.52	0.52	0.52	1614
weighted avg	0.52	0.53	0.53	1614

Doğruluk (*accuracy*) yaklaşık **%53**, makro ortalama F1 **0.52** civarındadır.

Konfüzyon matrisi incelendiğinde, özellikle A ve B segmentleri arasında karışmaların olduğu gözlenmektedir. D segmenti en net ayrışan sınıftır (yüksek recall).

6 Sonuçların Analizi ve Yorumlanması

6.1 Model Başarısı

Rastgele orman (RandomForest) ile %53 doğruluk elde edilmiştir. Verisetinin karmaşık yapısı göz önüne alındığında, performans orta seviyededir. Bazı segmentlerde (özellikle B) modelin yanılma payı yüksektir.

6.2 Hata Analizi

- A-B segmentleri arasında hatalı sınıflandırmalar gözlenir.
- D segmenti daha net tespit edilir (yüksek recall).

D segmentindeki müşterilerin profile daha tutarlı özellikler sergilediği düşünülebilir.

6.3 Uygulama Alanları ve Öneriler

- Pazarlama kampanyaları, bu segmente (ör. D) göre özelleştirildiğinde başarı artabilir.
- B segmentini iyileştirmek adına ek veri veya farklı özellikler (feature engineering) faydalı olabilir.

7 Sonuçların Raporlanması ve Sunumu

7.1 Proje Metrikleri

1. **Model Seçimi (30 puan)**: RandomForest seçildi, GridSearchCV ile parametreleri belirlendi, doğruluk %53 civarında.
2. **Sonuçların Analizi (35 puan)**: Confusion matrix, precision/recall, F1-skor incelendi. Segmentler arası karışmalar analiz edildi, "B" segmentinde zorluk görüldü.
3. **Sonuçların Raporlanması (35 puan)**: 6 adet grafik, tablo ve açıklamalar. Modelin doğruluğu, hata analizi, uygulama önerileri verilmiştir.

my_submission.csv dosyası, test veri setindeki müşterilerin öngörülen (A, B, C, D) segmentlerini içerir. Firma bu sonuçlardan yararlanarak özelleştirilmiş pazarlama stratejileri geliştirebilir.

8 Genel Değerlendirme

- Modelin doğruluğu %53'tür, segmentlerin ayrıştırılma güçlüğü söz konusudur.
- Diğer modeller (XGBoost, LightGBM, vb.) veya ek veriyle performans artabilir.
- Eksik değer stratejilerinde ve feature engineering aşamalarında yapılacak ek çalışmalarla sonuçlar iyileştirilebilir.

Kaynaklar

- Kaggle "Customer Segmentation" Dataset: `https://www.kaggle.com/datasets/vetrirah/customer`
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
- Scikit-Learn Dokümantasyonu: `https://scikit-learn.org/stable/`

Proje Ekibi:
Can Çebi (cancebi21@istanbularel.edu.tr)
Oğuzhan Cem Yücel (oguzhancemyucel21@istanbularel.edu.tr)

Teşekkürler: Bu rapor, Veri Madenciliği dersi kapsamında değerli katkı ve yönlendirmeleri için Sibel Birtane Akar hocamıza ve bölüm arkadaşlarımıza ithafen hazırlanmıştır.