

Hepsiemlak - Oğuzhan Gündüz

1. Soru — Yönetim Raporu için Veri Hazırlığı

Amaç

- İlan verisinden **yönetim raporu** üretmek istiyoruz.
- Bunun için veriyi raporlamaya uygun, temiz ve tekilleştirilmiş bir forma getirmeliyiz.

Final tablo ihtiyacı

- İlanların lokasyon vb. alanlarını doğru eşleştirebilmek için önce final bir tablo kurgulamalıyız.
- İşe LocationMapping ve Listings tablolarını birleştirerek başladım. Ardından yaptığım kontrollerde m², fiyat ve PPsqm alanlarında makul aralıkların dışına taşıyan değerler olduğunu gördüm; bu da bizi outlier temizliği ve DQ kuralları tanımlamaya yöneltti.

DQ (Data Quality) kontrolleri

Outlier temizliği — yüzdelik (percentile) tabanlı trim

- Outlier'ları Ay × Şehir × Kategori segmentinde p01–p99 yüzdeliklerine göre temizliyoruz.
- Segmentte örnek azsa (n < 10), Şehir × Kategori → Kategori → Global zinciriyle kademeli fallback uyguluyoruz.
- p01'in altını ve p99'un üstünü trimliyoruz (kayıtları dışlıyoruz).

Neden böyle?

Uç değerler TL/m² ve maliyet KPI'larını yanıltıyor. Segment bazlı p01–p99, **yerel piyasa seviyesini** yakalarken, düşük hacimde fallback ile güvenliği koruyor. Böylece KPI'lar **gerçeğe daha yakın** ve **karar verilebilir** oluyor.

Trim sonrası iş kuralları (guard-rails)

Trim tek başına yetmeyen yerlerde, **Satılık** ve **Kiralık** için ayrı bantlar uyguluyoruz ve **AND** ile sıkılaştırıyoruz:

Satılık (Sale) bantları

- ListingSqm ∈ [20, 1000]
- Price_TL ∈ [100.000, 500.000.000]
- PricePerSqm_TL ∈ [1.000, 500.000] (TL/m²)

Kiralık (Rent) bantları

- ListingSqm ∈ [15, 1000]
- Price_TL ∈ [1.000, 5.000.000]
- PricePerSqm_TL ∈ [50, 50.000] (TL/m²)

Ek kurallar

- p01/p99'u hesaplanmamış (p01 NULL) kayıtları dışlıyoruz.
- Tüm DQ bayraklı kayıtları atıyoruz: NOT is_low_ppsqm , NOT is_high_ppsqm , NOT is_bad_sqm , NOT is_currency_suspect .

Çıktı (strict çekirdek set): ww_listings_reporting_clean_strict — iş kuralları + DQ filtreleri uygulanmış, outlier'sız raporlama tabanı.

Coğrafya (GEO) katmanı

- "Strict" görünümümüzden yola çıkıp il/ilçe/mahalle alanlarını üretiyoruz.
- Türkiye ve **Kıbrıs** için uygun adres formatlarını ayrı ele alıyoruz.
- **District** seviyesinde özet KPI'lar hesaplıyoruz:
 - Satılık/kiralık ortalama TL/m²
 - Satılık/kiralık ortalama fiyat
 - Gözlem sayısı (adet)
- Çıktı: vw_geo_district coğrafi segmentasyon ve bölgesel kıyaslar için temel kaynak.

Neden "strict" bırakıyoruz?

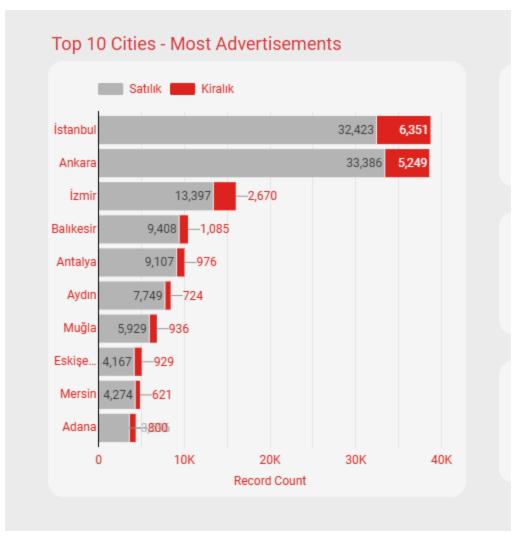
- Sunumda yanıltıcı yüksek/çok küçük değerlerin KPI'ları bozmasını istemiyoruz.
- "Strict" = outlier'ı at, DQ ve bant kurallarını katı uygula.

Final görünüm seti

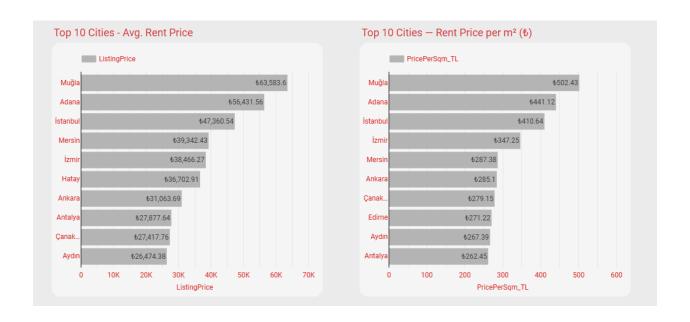
- vw_listings_reporting_clean_strict → Analiz için çekirdek veri seti (trim + DQ + bantlar).
- vw_geo_district → District bazlı fiyat ve TL/m² ortalamaları + adetler.

Bu hazırlık ve kaynak **view**'ları oluşturduktan sonra, veriyi doğrudan **Looker Studio**'ya bağladık ve raporlama aracı olarak Looker'ı kullandık. Soruda istenen her şey ve fazlasını **dashboard**'ımızda görebilirsiniz:

<u>hepsiemlak - Oğuzhan Gündüz - Listed Advert Analysis</u>

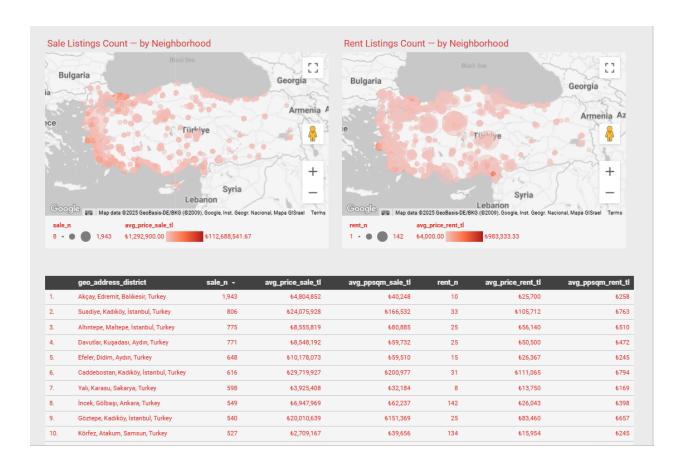






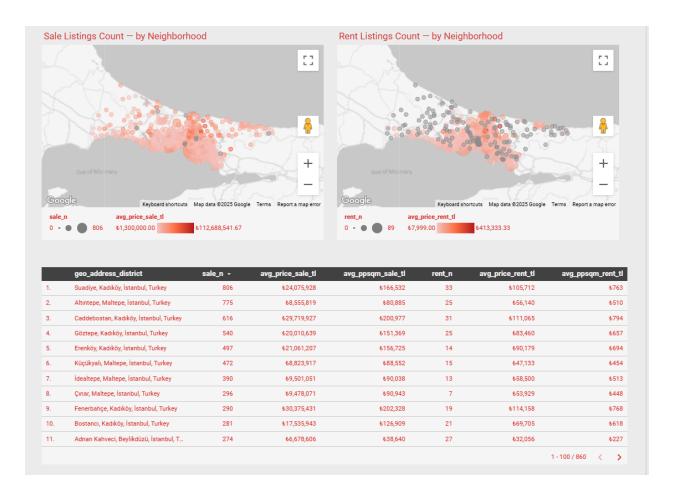
<u>hepsiemlak - Oğuzhan Gündüz - Listing Category & Location</u>

City	Record 🔻	Avg. Listing	Avg. Listing	Avg. PPSQ	Avg. Listing	City	Record ▼	Avg. Listing	Avg. Listing	Avg. PPSQ	Avg. Listing
Ankara	33,386	\$ 5,584,658.18	146.32	38,167.31	103.75	İstanbul	6,351	₺47,360.54	113.4	417.63	
İstanbul	32,423	₺11,840,325.31	137.96	85,823.7	102.77	Ankara	5,249	₺31,063.69	117.55	264.26	
İzmir	13,397	₺7,519,963.55	140.87	53,382.98	102.06	İzmir	2,670	\$38,466.27	108.81	353.51	
Balıkesir	9,408	€6,293,468.74	133.77	47,047.47	102.36	Balıkesir	1,085	€22,658.2	95.8	236.52	
Antalya	9,107	€6,865,111.22	131.17	52,338.57	103.46	Antalya	976	₺27,877.64	115.11	242.19	
Aydın	7,749	₺8,492,505.86	162.67	52,208.19	108.73	Muğla	936	₺ 63,583.6	120.9	525.91	
Muğla	5,929	12,990,862.5	151.24	85,894.05	103.35	Eskişehir	929	₺17,198.92	90.13	190.82	
Mersin	4,274	€4,278,397.93	142.43	30,038.15	102.89	Adana	800	₺ 56,431.56	116.47	484.5	
Eskişehir	4,167	₺3,775,642.43	123.76	30,507.12	101.98	Aydın	724	€26,474.38	108.47	244.07	
Kayseri	3,833	\$ 3,795,823.89	170.8	22,223.47	100.38	Mersin	621	₺39,342.43	127.78	307.88	
Samsiin	3 635	£3 491 378 23	119 12	29 309 97 1 - 73	97 52 / 73 〈 〉	Tekirdaň	574	\$23.266.QQ	102 41	227 18 1 - 61	/61 <



hepsiemlak - Oğuzhan Gündüz -Istanbul Breakdown

County	Record 🕶	Avg. Listing	Avg. Listing	Avg. PPSQ	Avg. Listing	County	Record +	Avg. Listing	Avg. Listing	Avg. PPSQ	Avg. Listi	ng
Kadıköy	4,119	621,384,504.3	136.55	156,602.18	105.01	Şişli	631	€43,000.78	97.5	441.02		9
Maltepe	2,671	\$8,527,394.61	109.23	78,071.55	106.01	Kağıtha	598	€32,512.45	87.06	373.47		9
Küçükç	1,514	₺5,565,862.41	116.75	47,672.94	102.75	Beşiktaş	423	€78,586.41	125.4	626.66		9
Esenyurt	1,468	€3,429,579	118.29	28,992.43	101.99	Sariyer	395	€90,461.14	170.19	531.53		9
Beylikd	1,252	66,612,817.34	159.91	41,353.44	103.42	Kadıköy	387	€67,907.07	113.48	598.42		9
Bahçeli	1,250	€7,269,480	121.42	59,868.17	106.5	Beyoğlu	320	645,654.84	83.5	546.76		9
Kağıtha	1,248	€6,240,062.18	103.41	60,343.25	101.56	Üsküdar	314	€53,295.38	118.68	449.09		9
Şişli	1,184	€10,350,811.66	117.6	88,014.2	100.54	Esenyurt	294	€19,654.25	101.18	194.25		9
Üsküdar	1,166	₺15,938,587.91	148.71	107,179.52	109.67	Fatih	233	€25,238.63	88.52	285.13		9
Beşiktaş	970	€29,466,003.09	171.4	171,915.74	102.18	Küçükç	209	€31,385.17	101.18	310.2		9
Fatih	941	£5 ∆11 R12 Q6	107 07	50 544 06 1 - 39	102 72	Maltene	202	\$.42 RN2 4R	104 24	410 62 1 - 38	/38 <	9



***Hepsi ve daha fazlası için Dashboarda göz atabilirsiniz.

2. Soru — Retention (Aşama 1: Mevcut Durum Analizi)

Amaç

- Retention tarafında "kimler yeniliyor / yenilemiyor?" sorusuna giden **temel** analitik zemini kurmak istiyoruz.
- Bunun için önce veri yapısını doğru okuyup, iş kararının verildiği seviyeyi (kontrat) netleştiriyoruz.

Veri yapısını nasıl okuduk?

- Yıllık kontrat paterni: Paketlerin start-end tarihlerini incelediğimde, büyük ölçüde 12 aylık bir döngü gördüm.
- Aylık satırlar vs. yıllık alanlar: period bazlı satırlarda performans metrikleri (pageview, listings, conversion) aylık değişiyor; buna karşılık annual niteliğindeki alanlar (ör. annual fee, bazı statüler) her period'da aynen tekrarlanıyor.
- Retention tablosu = join'lenmiş kaynaklar: Bu tablo, farklı tablolardan gelen alanları birleştiriyor gibi; özellikle fee ve bazı "abonelik" statüleri aylık değil, yıllık. O yüzden "aylık satırda tekrarlanan yıllık alanı" aylık analizde birebir kullanmak yanıltıyor.

Buradan çıkan karar (yaklaşımımız: simple)

- Analizi kontrat seviyesine indirgedik: FirmKey + FirmPackageKey + end_dt = tek kontrat.
- Bu sayede:
 - Yıllık alanları (örn. annual fee) tekilleştirip kontrat satırına koyuyoruz.
 - Aylık performansları (pv, listings, conversion) yıl bazında topluyoruz (Σ aylık).
 - Bu toplamlar üzerinden oran/KPI üretiyoruz:

```
conv_rate_year = Σconv / Σpv
```

- pv_per_listing_year = Σpv / Σlistings
- cost_per_view_year = annual_fee / Σpv
- cost_per_listing_year = annual_fee / Σlistings
- cost_per_conversion_year = annual_fee / Σconv
- Bu yapıya "simple" adını verdik: yalın, tekrarsız, karar anına odaklı.

Etiket (gerçekleşen) nasıl belirlendi?

- Actual retention label'ı, kontrat bitiş ayındaki statüye göre aldık:
 - o label_contract = 1 → Yeniledi

- label_contract = 0 → Yenilemedi
- (Bitiş ayında etiket görünmüyorsa son görülen etiketi yedek kabul ettik.)
- Böylece "gerçekleşen" (actual) retention, kontrat bazında saf biçimde ölçülür hale geldi.

Neden kontrat-level doğru?

- Aylık satırlarda aynı etiketin aylar boyunca tekrar etmesi ve annual alanların her ay kopyalanması, hem KPI'ları hem de (ileride) ML'i yanlı yapıyor.
- Kontrat-level'e inince:
 - Aşırı ağırlık (aynı firmanın aylarca tekrar yazılması) kalkıyor.
 - KPI'lar "yıllık ücret ↔ yıllık performans" denkleminde adil hesaplanıyor.
 - "Yeniledi mi/yenilemedi mi?" kararı zaten bitiş döneminde verildiği için, analiz doğru zamana bağlanıyor.

Çıktılar (mevcut durum KPI'ları)

- Genel yenileme oranı (labeled kontratlar üzerinden).
- **Paket bazında** yenileme oranı + yıllık maliyet/performans KPI'ları (cost per view/listing/conversion, conversion rate, views per listing).
- Aylık trend (paket x dönem): pv, listings, conv ve türetilmiş oranların zaman içindeki seyri.
- Bu KPI'ları dashboard'da hem **genel** hem **paket kırılımında** izliyoruz.

tabii—aynı madde başlıklarını **gerçek view adlarıyla** yazıyorum:

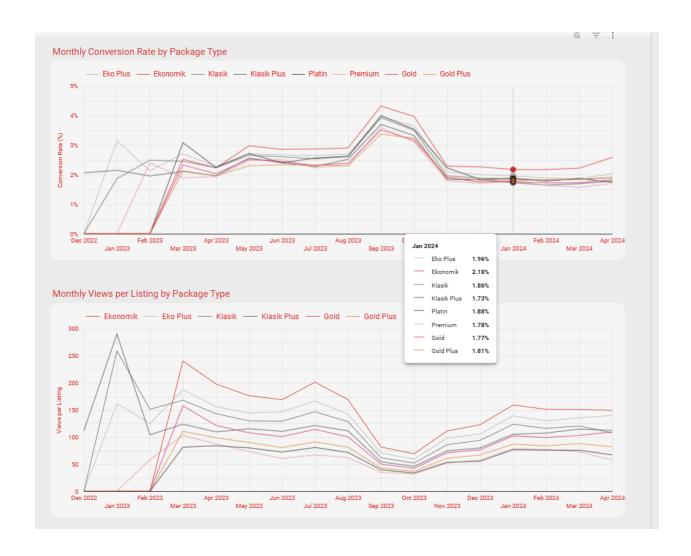
- Kontrat listesi (tek satır/kontrat) vw_dash_contract_list_simple
 (Yıllık performans toplamları + annual fee + actual label + türetilmiş KPI'lar)
- Paket bazında KPI özetleri (weighted, ratio-of-sums) vw_dash_kpi_by_package_simple
- Genel KPI (tüm paketler, toplu) vw_dash_kpi_total_simple
- Paket × dönem (aylık) trend vw_dash_package_monthly_simple

Not: Bu simple (kontrat-level) zeminle ilerleyip ML'e geçtiğimizde metrikler belirgin şekilde iyileşti;

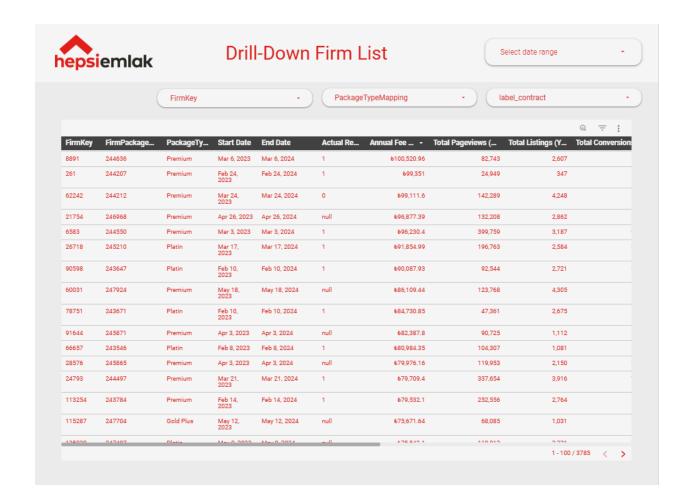
hepsiemlak - Oğuzhan Gündüz - Retention & KPI's by Package Type



<u>hepsiemlak - Oğuzhan Gündüz - Package Type x Period Trend</u>



<u>hepsiemlak - Oğuzhan Gündüz - Drill-Down Firm List</u>



2. Soru — Retention (Aşama 2: ML Katmanı)

Amaç

- "Kimler yeniler/yenilemez?" sorusunu kontrat seviyesinde tahminlemek istiyoruz.
- Aylık tekrar ve etiket kopyaları yüzünden oluşan yanlılığı kaldırmak için simple (contract-level) veri modelimizi kullanıyoruz.

Veri seti ve etiket (actual)

- Her satır tek kontrat: FirmKey + FirmPackageKey + end_dt .
- Yıllık alanlar tekilleştirildi (ör. annual_fee), aylık metrikler yıl toplamı olarak toplandı:

sum_pv, sum_listings, sum_conv → buradan oranlar/KPI:

pv_per_listing_year, conv_rate_year, cost_per_view_year, cost_per_listing_year, cost_per_conversion_year .

Etiket (label_contract): Bitiş ayındaki statü → 1=Yeniledi , 0=Yenilemedi ,
 NULL=Bilinmiyor (NULL'lar eğitime girmez, sadece skorlanır).

Kaynak (contract-level analitik tabanı): vw_dash_contract_list_simple

Train / Test stratejisi (zaman temelli)

- Son dönemde tek sınıf riski gördük; bu yüzden TEST'i "son iki sınıflı ay" olarak seçiyoruz.
- Uygulamada: end_period_ym bazında pozitif ve negatif içeren en güncel ay TEST, diğer etiketli kontratlar TRAIN.
- Bu, canlı hayattaki kullanım senaryosunu (geçmişle eğit, en yeni dönemde ölç) doğru yansıtıyor.

Not: vw_dash_contract_list_simple içinden etiketi bilinen kontratları alıp zaman bölmesi yapıyoruz. Eğitim ve değerlendirme sorgularını bu mantığa göre çalıştırdık.

Model

BigQuery ML – Logistic Regression

Model adi: hepsiemlak-470514.hepsiemlak_case.retention_lr_simple

Ayarlar: sınıf dengesizliği için AUTO_CLASS_WEIGHTS=TRUE (diğer hyperparametreler varsayılan).

• Öznitelikler (feature set):

annual_fee, sum_pv, sum_listings, sum_conv, pv_per_listing_year, conv_rate_year, PackageTypeMapping (+ ihtiyaç duyduklarımız).

Değerlendirme ve eşik seçimi

• ML.EVALUATE Çıktısı (varsayılan **0.50** eşik) bize genel bir tablo verdi; ancak **operasyonel karar** için **eşik taraması** yaptık.

- **0.01 adımlı** grid ile **F1**'i taradık ve **en iyi eşik = 0.43**'ü bulduk:
 - o F1≈ 0.80
 - Precision ≈ 0.707 (tahmin "yeniler" dediğimizin ~%71'i gerçekten yeniliyor)
 - Recall ≈ 0.922 (yenileyenlerin %92'sini yakalıyoruz)

Özet: Eşiği 0.43'e çekince kurtarma/upsell için daha uygun bir denge yakaladık (yüksek recall + kabul edilebilir precision).

Skorlama ve segmentasyon

- Kimi skorladık? Etiketi bilinmeyen (veya güncel/pipeline) kontratlar.
- Nasıl sunduk?
 - Decile dağılımı (1=düşük olasılık, 10=yüksek) ve ort. olasılık →
 vw_ml_decile_dist_simple
 - Özet (end_dt × paket × risk band): adet, medyan/ortalama olasılık →
 vw_ml_summary_simple
 - Operasyon kuyruğu (High Risk odaklı):

```
Expected Fee at Risk = (1 − renew_prob) × annual_fee , gün bazında sıralı liste → vw_ml_ops_queue_simple
```

Risk Band tanımı (raporda kullandığımız):

• High: Decile 1-3 (düşük olasılık)

• Medium: 4-7

Low: 8-10 (yüksek olasılık)

Neden bu yaklaşım daha iyi sonuç verdi?

- Aylık satırlardaki etiket kopyası ve aynı firmanın tekrarları kaldırıldı → veri sızıntısı ve ağırlık çarpılması yok.
- Karar seviyesini kontrat olarak netledik → model gerçekten verilmekte olan kararı öğreniyor.

Bu sayede önceki denemelere göre (aylık satır bazlı) metrikler belirgin şekilde
 iyileşti; F1=0.80'e kadar çıktık.

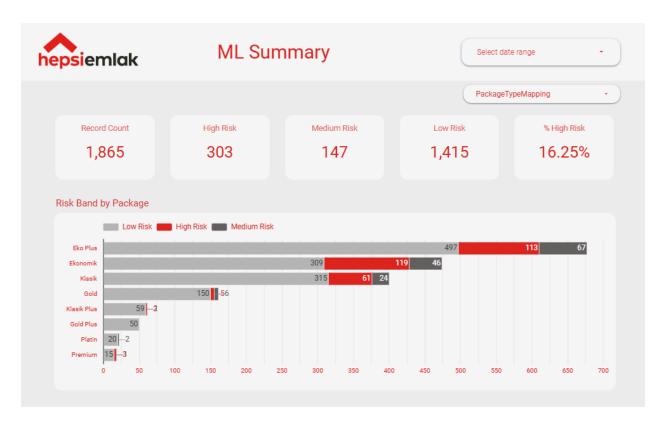
Dashboard entegrasyonu (kullanım)

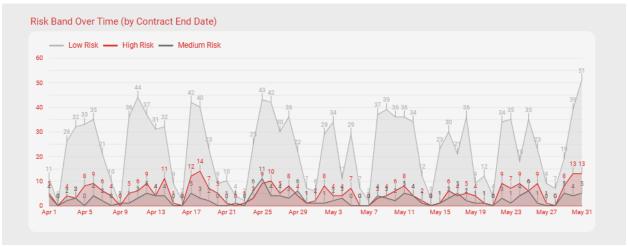
- Distribution sayfasında decile histogramı ve ortalama olasılık (kaynak: vw_ml_decile_dist_simple).
- Ops Queue (High Risk) sayfasında günlük iş listesi ve Expected Fee at Risk ile önceliklendirme (kaynak: ww_ml_ops_queue_simple).
- ML Summary sayfasında end_dt × paket × risk band adet/olasılık özetleri (kaynak: vw_ml_summary_simple).
- Detaylı Firma Listesi tablosunda analitik (actual) + skor (predicted) alanları birlikte gösteriyoruz.

Kısa sonuç:

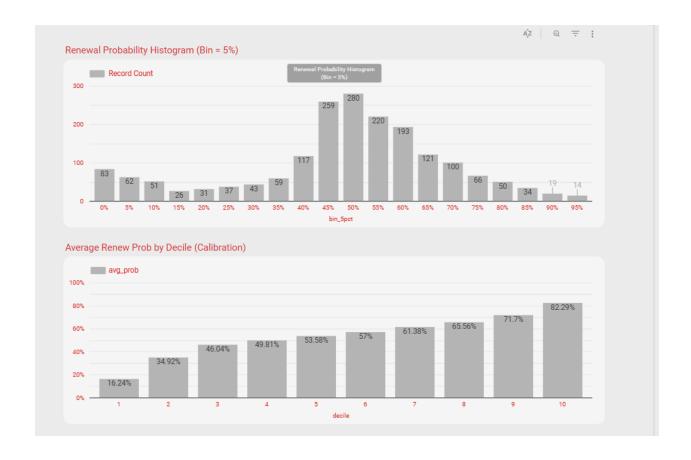
Simple (contract-level) veri modeli + zaman temelli test + eşik optimizasyonu (0.43) ile, operasyonu besleyen yüksek recall'lı ve iş açısından anlaşılır bir tahmin katmanı kurduk. Üretim kullanımı için gerekli olan tüm view'lar yukarıda.

hepsiemlak - Oğuzhan Gündüz - ML Summary

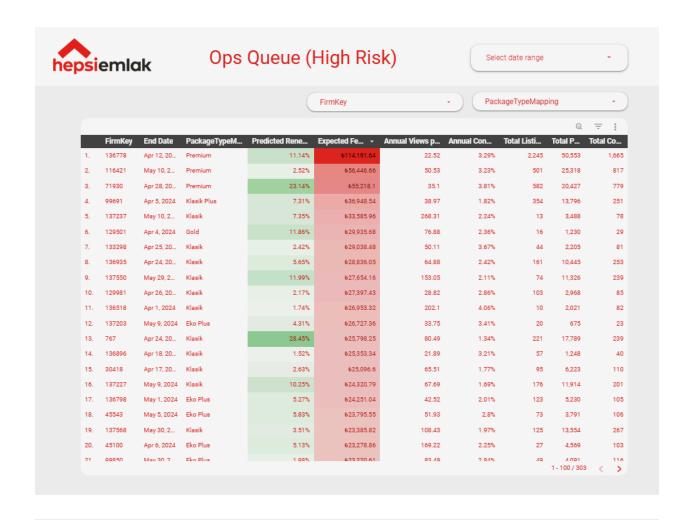




<u>hepsiemlak - Oğuzhan Gündüz - Distribution</u>



<u>hepsiemlak - Oğuzhan Gündüz - Ops Queue (High Risk)</u>



Son Söz: Dar bir zamanda ciddi bir efor harcayıp kendimi iyi anlamda challange ettiğim bir deneyim oldu. Uykusuz ve yorgun günler geçirdim fakat sonuçta elde ettiğim çalışma içime çok sindi.

Dolayısı ile ben de sizlere teşekkür ederim. bilmediğim noktada araştırıp, kendimi geliştirmek zorunda kaldığım bir çalışma oldu. günün sonunda kazançlıyım.

Sevgi ve Saygılarımla:.

Oğuzhan Gündüz.