Model için Veri Hazırlama: (EDA) | Enes Fehmi Manan

Öne Çıkan Maddeler

1. Problem Tanımı ve Senaryo

1. **Gerçekçi Senaryo:** Bankacılık sektöründe **batık kredi** oranlarının artması.
2. **Problem:** Son 3 ayda yükselen **batık kredi verme oranını** düşürmek.
3. **İş Süreci:** Problemin **üst yönetimden** gelmesi ve **business** ekipleriyle iletişim.
4. **İlk Yaklaşım: Neden soruları** sormak (oran neden yükseldi?).
5. **Veri Bilimcinin Rolü:** Sadece model kurmak değil, problemi anlamak ve **dataset oluşturmak**.

2. Machine Learning Öncesi Adımlar

1. **Drift Kavramı:** Model, data veya popülasyon **drift'i** olası bir sebep olarak incelendi.
2. **Threshold/Cutoff:** Mevcut kurallardaki **eşik değerlerinin** yanlış ayarlanmış olabileceği.
3. **Alternatif Çözümler:** Machine learning'e girmeden basit **kural setleri** veya **karar ağaçları** ile çözüm arayışı.
4. **Maliyet Analizi: Machine Learning** projesinin zaman, kaynak ve **deployment** maliyetlerinin düşünülmesi.

3. Veri Toplama ve Hazırlık

1. **Sıfırdan Veri Seti: Kaggle** gibi hazır veriler yerine gerçek bir **database**'den veri çekme zorluğu.
2. **Target Tanımı: Batık krediyi** tanımlamak için **90 Plus** (90+ gün gecikme) metriğinin kullanılması.
3. **Periyot Belirleme: Train ve test** periyotlarının doğru seçilmesi ve **data leak** riskinin önlenmesi.
4. **Domain Bilgisi: Feature** seçiminde sektör ve iş bilgisi kritik öneme sahiptir.
5. **Feature Seçimi: Binlerce kolon** arasından **analitik yaklaşımla** ve **business** desteğiyle doğru değişkenleri bulmak.
6. **Veri Beslenmesi:** Seçilen kolonların **train/test periyodunda** hala güncel veriyle beslenip beslenmediğinin kontrolü.

4. Keşifsel Veri Analizi (EDA)

1. **EDA'nın Amacı:** Veriyi anlamak, dağılımları görmek ve anormallikleri tespit etmek.
2. **Veri Tipleri:** Kolonların **doğru veri tipinde** (numerical, categorical) olduğundan emin olmak.
3. **Betimsel İstatistikler:** Ortalama, medyan, standart sapma gibi temel istatistiklere bakmak.
4. **Eksik Değer Analizi:** Yüksek oranda **eksik değere** sahip kolonları tespit etmek ve strateji belirlemek.
5. **Aykırı Değer (Outlier) Tespiti:** Aykırı değerlerin tespiti ve business ile konuşarak nasıl ele alınacağına karar vermek.
6. **Dağılım Analizi: Train ve test setlerinin dağılımlarının** birbiriyle uyumlu olup olmadığını kontrol etmek.
7. **Olasılık Yoğunluk Fonksiyonları:** Sürekli değişkenlerin dağılımını görselleştirmek.
8. **Kategorik Değişken Analizi:** Kategorik verilerin sınıflarını ve dağılımlarını incelemek.
9. **Target'a Göre Analiz:** Değişkenlerin dağılımını **target** (batık/sağlam) kırılımında inceleyerek ayırt edici özellikler aramak.
10. **Korelasyon Analizi:** Değişkenler arası ilişkileri ve **multicollinearity** riskini tespit etmek.
11. **Otomatik EDA Kütüphaneleri: Pandas Profiling**, **Sweetviz** gibi araçlarla EDA sürecini hızlandırmak.

5. Baseline Model ve Modelleme

1. **Baseline Model:** En basit haliyle, minimal ön işleme ile kurulan ilk model.
2. **Lojistik Regresyon:** Açıklanabilirliği yüksek olduğu için sıkça kullanılan bir **baseline** modeli.
3. **Scikit-learn Pipeline:** Ön işleme ve modelleme adımlarını bir arada tutan verimli bir yapı.
4. **Dengesiz Veri Seti:** Target dağılımı dengesiz olduğu için **stratify** parametresiyle veri ayırma ve **class_weight** ile modeli eğitme.
5. **Ön İşleme Adımları: StandardScaler** ile normalizasyon, **SimpleImputer** ile eksik değer doldurma, **One-Hot Encoding** ile kategorik veri dönüşümü.
6. **Model Değerlendirme Metrikleri: AUC**, **Precision**, **Recall** ve **F1-Score**.
7. **Confusion Matrix:** Modelin **True Positive/Negative** ve **False Positive/Negative** tahminlerini detaylı görmek.
8. **Overfitting Kontrolü: Train** ve **validasyon** skorları arasındaki farkı gözlemlemek.

6. Model İyileştirme ve Sonraki Adımlar

1. **İteratif Süreç:** Modellemenin tek seferlik değil, sürekli bir iyileştirme döngüsü olduğu.
2. **Model Değişimi: Baseline** sonrası **LightGBM**, **XGBoost** gibi daha güçlü algoritmaları denemek.
3. **Feature Engineering:** Mevcut değişkenlerden yeni ve daha anlamlı **feature'lar** türetmek.
4. **Feature Setini Genişletmek:** Daha fazla veri kaynağından yeni değişkenler eklemek.
5. **Feature Reduction:** Model performansını düşürmeden en anlamlı **feature'ları** seçerek modeli basitleştirmek.
6. **SHAP ve Feature Importance:** Modelin hangi değişkenlere daha fazla önem verdiğini anlamak.
7. **Unfair Değişkenler:** Cinsiyet, yaş gibi **etik dışı** veya regülasyona aykırı olabilecek değişkenleri final modelden çıkarmak.
8. **Regülasyonlar:** Bankacılık gibi sektörlerde **BDDK** ve **Basel** gibi regülasyonlara uyum zorunluluğu.

7. Araçlar ve Kavramlar

1. **MLflow:** Model deneylerini, parametreleri, metrikleri ve **artifact'ları** (modelleri) takip etmek için kullanılan bir araç.
2. **Deney Takibi:** Hangi değişikliğin skoru nasıl etkilediğini sistematik olarak kaydetmenin önemi.
3. **Sampling Yöntemleri: Undersampling** ve **SMOTE** gibi tekniklerin teoride var olduğu ancak pratikte dikkatli kullanılması gerektiği.
4. **PSI (Population Stability Index):** Feature'ların zaman içindeki dağılım istikrarını ölçen bir metrik.
5. **Kaggle:** Gerçekçi veri setleri ve yarışmalarla pratik yapmak için önemli bir platform.

8. Kariyer ve Sektör

1. **Portfolyo Önemi: Uçtan uca projeler** ve **hackathon/yarışma** deneyimlerinin işe alımda öne çıkarıcı olduğu.
2. **Temel Bilgiler:** İstatistik, **cross-validation** gibi temel kavramlara hakimiyetin önemi.
3. **Sektör Daralması:** Alana olan yoğun ilgi nedeniyle öne çıkmak için **derinlemesine projeler** yapmanın gerekliliği.

Problem Tanımı ve Yaklaşım

Bu yayında, bir bankada çalışan veri bilimcinin karşılaştığı **gerçekçi bir senaryo** üzerinden **machine learning için veri hazırlama** süreci simüle edilmektedir. Temel problem, son üç ayda artan **batık kredi verme oranının** düşürülmesidir. İlk adımda, probleme **neden soruları** sorularak yaklaşılmış ve olası sebepler (**drift**, **data kayması**, **popülasyon kayması**, **yanlış threshold**) tartışılmıştır. Machine learning'e geçmeden önce, problemi basit **kural setleri** veya **karar ağaçları** ile çözme olasılığı değerlendirilmiş, ancak sorunun karmaşıklığı nedeniyle **machine learning** yaklaşımına karar verilmiştir. Bu noktada, sıfırdan bir **dataset oluşturma** gerekliliği vurgulanarak sürecin pratik zorluklarına dikkat çekilmiştir.

EDA ve Baseline Model

Veri hazırlama sürecinin merkezinde, **target tanımı** (bu senaryoda **90 Plus** gecikmeye düşen müşteriler) ve uygun **train-test periyotlarının** belirlenmesi yer almaktadır. Gerçek dünyadaki gibi büyük bir **database** içerisinden, **domain bilgisi** ve **business ekipleriyle** iş birliği içinde potansiyel **feature'ların** seçilmesi kritik öneme sahiptir. Seçilen veriler üzerinde **Keşifsel Veri Analizi (EDA)** uygulanmış; **veri tipleri**, **eksik değerler**, **aykırı değerler**, **dağılımlar** ve **korelasyonlar** incelenmiştir. **Dengesiz veri setleri** için **sampling** ve **ağırlıklandırma** gibi teknikler tartışılmış, **train-test dağılım uyumu** ve **kategorik değişkenlerin** target'a göre dağılımı görselleştirilmiştir. Bu analizlerin ardından, **scikit-learn pipeline** yapısı kullanılarak minimal ön işleme ile basit bir **baseline model (Lojistik Regresyon)** oluşturulmuş ve ilk sonuçlar **Confusion Matrix** ve **AUC** metrikleri ile değerlendirilmiştir.

Model Geliştirme ve Deney Takibi

İlk **baseline modelin** sonuçları, özellikle yüksek **False Positive** oranıyla, yetersiz bulunmuştur. Bu noktadan sonra izlenecek adımlar olarak **model değiştirme** (**LightGBM**, **XGBoost** gibi daha güçlü modelleri denemek), **feature setini genişletmek**, **feature engineering** yapmak ve **etik dışı/regülasyona aykırı (unfair)** feature'ları modelden çıkarmak gibi stratejiler önerilmiştir. Model geliştirme sürecindeki tüm bu adımların ve deneylerin takibi için **MLflow** gibi araçların önemi vurgulanmıştır. Sonuç olarak, veri hazırlama ve modelleme sürecinin tek seferlik değil, **iteratif** bir süreç olduğu ve sürekli olarak **deney yapma**, **sonuçları analiz etme** ve **business hedefleriyle** uyumlu hale getirme çabası gerektirdiği belirtilmiştir.