Final Proje Detayları | Enes Fehmi Manan

Anlatan: Enes Fehmi Manan

Moderatör: Özge Usta

---

1. Giriş ve Kaynaklar

Enes Fehmi Manan:
Herkes hoş geldi diyelim. Daha önceki yayından hatırlarsınız, ben orada eğitim notebook'unu göstermiştim. Hatta o notebook'u analiz klasörünün altına gönderdim, oradan ulaşabilirsiniz. Kabul edilen notebook'ları da zaman içerisinde biraz düzenleyip atacağım.

Buraya baktığınız zaman repo bir tık büyük ve karmaşık gelebilir. Buradaki seçtiğim yarışma ve dataset de aslında bir tık karmaşık. Sizin bu kadar karmaşık bir şeye baştan gitmenize gerek yok, bunu baştan söyleyeyim ki gözünüz korkmasın.

Buradaki projenin zorluk seviyesi, Cemil hocamın da dediği gibi, aslında sizin seçeceğiniz dataset ile ve sizin projeye ne kadar emek vereceğinizle doğru orantılı. O kısımda sizi biraz özgür bıraktık. Tamamen kendi seçeceğiniz doğrultuda, dataset ile çalışacaksınız ve işin sonunda ne kadar emek verirseniz projede o kadar derinleşirsiniz diyelim.

Şimdi yavaştan proje dökümanına geçeyim. Biz bu bootcamp kapsamında nasıl bir şey istiyoruz? Bu dökümanı sizinle paylaşacağım.

2. Proje Gereksinimleri

Enes Fehmi Manan:
Machine Learning Bootcamp'in uçtan uca final projesi nedir? İstediğimiz şey şu: **Kendi belirlediğiniz bir sektördeki probleme, uçtan uca makine öğrenmesi ile bir çözüm geliştirmenizi ve çözümün her adımını özgürce dökümante etmenizi istiyoruz.**

Buradaki önemli noktalardan biri, sizi özgür bırakmamız. Sektörü de problemi de siz seçeceksiniz. Datayı da siz bulacaksınız. Buna ne kadar ön işleme, ne kadar Feature Engineering yapacağınızı, hatta Business Case'i kendiniz belirleyip ona göre bir şeyler dizayn etmenizi isteyeceğiz.

Proje Gereksinimleri:

1) Sektör ve Problem Seçimi:
Benim gösterdiğim örneğe bakarsak; Bankacılık sektörü var. Problem ne? Kredi riskinde, "ödeyecek" deyip kredi verdiklerimin gecikmeye düşmesi ve bankanın para kaybetmesi. Buradaki amaç, bir makine öğrenmesi modeliyle bunu önceden tahmin etmek.
Bu başka bir şey olabilir; bir fraud (dolandırıcılık) modeli olabilir, kampanya tarafında bir şeyler olabilir, e-ticaret sektöründe recommendation (öneri) sistemi olabilir, churn (müşteri terk) analizi olabilir, oyun tarafında segmentasyon veya LTV (Life Time Value) prediction olabilir.

Kendi keyfinize göre bir problem bulabilirsiniz. Burada önemli kriter; **hoşunuza gidecek, üzerinde vakit ayırırken sıkılmayacağınız bir şey seçmeniz.** Veya hedeflediğiniz bir sektöre yönelik portfolyo projesi de olabilir.

2) Dataset Seçimi:

Sektörü ve problemi seçtik. Buna uygun dataset'i nereden bulacağız?

Kaggle'da geçmiş veya devam eden yarışmalardan bir dataset bulmanızı tavsiye ediyoruz.

Bunun sebebi ne?

- Hacim olarak hatırı sayılır seviyede oluyorlar.
- Description (Açıklama) kısımları ve insanların çözümlerini paylaştığı "Write-up"lar oluyor. Gelip oradaki notebook'ları, çözümleri ve yaklaşımları görebiliyorsunuz. Bu inanılmaz ufuk açıcı bir şey.
  Özellikle düz bir Kaggle dataset sekmesinden data seçmektense, Kaggle'da bir **yarışma datası** bulup onu kullanmayı şiddetle tavsiye ediyoruz.

Dataset Kriterleri:

- **Format:** Tabular (Tablo) data üzerinden gidiyoruz. CSV, Parquet veya Excel formatında olmalı. Görsel veya ses gibi şeyler olmamalı.
- **Sentetik Olmamalı:** Tercihen sentetik bir dataset olmasın. Dataset description'larına bakabilirsiniz, sentetikse belirtiyorlar. Sentetik data kendini belli ediyor; başarılar çok yüksek geliyor, çok atraksiyon (feature engineering vb.) yapamıyorsunuz. Portfolyo açısından da gerçek data daha önemlidir.
- **Playground Yarışmaları:** Kaggle'ın "Playground" serisi yarışmaları genellikle sentetik data içerir. Bunları seçmenizi tavsiye etmem. Gerçekten para ödüllü (geçmiş veya güncel) yarışmaları seçmenizi öneriyorum.
- **Data Boyutu:** En az **10.000 satır** olmalı. Küçük datada modelleme çok mantıklı olmuyor. En az 10k olsun ki bir feature engineering yapabilin, farklı kolonlar çıkarıp ekleyebilin. Tercihen 80-90 feature olsun ama korkuyorsanız daha azı da olabilir, sadece bu kriterlerin çok altına inmemeye çalışalım.

Veri kaynağı olarak sadece Kaggle olmak zorunda değil; UCI Machine Learning Repository, Google Dataset Search veya Hugging Face de olabilir. Ama Kaggle yarışma datası bir adım öndedir.

3. Repo ve Proje Yapısı

Enes Fehmi Manan:

Peki bu projenin bir Kaggle yarışmasından farkı nedir?

Kaggle yarışmasında her şey skordur. Lokal validasyonunuz iyiyse atarsınız ve skorunuz yükselir. Buradaki fark; birazcık daha **Business ve sistemsel kurguyu düşünmenizi** istiyoruz. İşin sonunda düzenli bir repo yapısı bekliyoruz.

Bir repo yapısı nasıl olabilir? (Örnek repo üzerinden anlatıyor):

- docs klasörü: Dökümanlar.
- models klasörü: Kaydedilen modeller.

- notebooks klasörü: Geliştirme aşamaları.
- src klasörü: API veya pipeline scriptlerinin bulunduğu yer.

Notebook Yapısı (Örnek Akış):

1. **EDA (Keşifçi Veri Analizi):** Problem tanımı, değişkenlere bakış, dağılımlar, korelasyonlar. LLM (Yapay Zeka) yardımı alabilirsiniz ama çıktıları kendiniz yorumlamalısınız. "Bu dağılım neden böyle?" diye düşünmelisiniz.
2. **Baseline Model:** En basit feature set ve basit bir modelle (örn: mean, mode veya basit bir algoritma) bir skor elde edin. Bu sizin referans noktanız olacak.
3. **Feature Engineering:** Burası yaratıcılığınızı konuşturacağınız yer. Dışarıdan data (altın fiyatı, faiz oranı vb.) ekleyebilirsiniz. Yaptığınız denemeleri notebook veya docs içinde açıklayın.
4. **Model Optimization:** Grid Search, Random Search veya Optuna gibi yöntemlerle belirlediğiniz uzayda modeli optimize edin.
5. **Model Evaluation:** Feature importance, SHAP değerleri ve Business yorumlaması.
6. **Final Pipeline:** Tüm adımları içeren, datayı alıp temizleyip modeli eğiten final notebook veya script.

Scripts (Source Klasörü):
Modeli eğittik, bir .pkl (pickle) formatında kaydettik. Bu modelden bir ekran (arayüz) bekliyoruz.

- Streamlit veya Flask kullanarak basit bir arayüz yapabilirsiniz.
- Örneğin: Kullanıcı değerleri girer, model arkada çalışır ve "Riskli/Risksiz" diye sonuç döner.
- Bunu Hugging Face Spaces, Render veya Streamlit Cloud üzerinde ücretsiz deploy edebilirsiniz.

4. README Dosyası ve İçeriği

Enes Fehmi Manan:
Projenin içerisinde bir README.md dosyası bekliyoruz. İçeriğinde neler olmalı?

- **Proje Başlığı:** Proje nedir?
- **Kapsam:** Zero2End Bootcamp final projesi kapsamında yapılmıştır.
- **Problem:** Çözdüğüm problem (örn: bankacılık kredi riski).
- **Çözüm:** Nasıl çözdüm? Hangi dataset'i kullandım?
- **Demo Linki:** Basit inference (tahmin) alabileceğimiz deploy linki.
- **Görseller:** Ekran görüntüleri veya mimariyi anlatan bir şema.
- **Metrikler:** Hangi metriği seçtiniz ve nasıl optimize ettiniz?
- **Teknolojiler & Kurulum:** Hangi kütüphaneler kullanıldı? Lokalimizde nasıl çalıştırırız?

5. Repoda Cevaplanması Gereken Kritik Sorular

Enes Fehmi Manan:

Repoda (Readme veya Notebook içinde) şu soruların cevabı yazılı olarak bulunmalı:

1. **Problem Tanımı:** Problemi net bir şekilde kendi cümlelerinizle tanımlayın.
2. **Baseline Süreci ve Skoru:** İlk denemeniz neydi, sonucu ne oldu?
3. **Feature Engineering Denemeleri:** Hangi feature'ları ürettiniz, hangileri işe yaradı? Bu kısım yaratıcılığınızı gösterir.
4. **Validasyon Şeması:** K-Fold mu, Time-based split mi? Neden bunu seçtiniz? Yanlış validasyon şeması vezir de eder rezil de eder. Bunu açıklayın.
5. **Ön İşleme Stratejisi:** Final feature setine nasıl karar verdiniz?
6. **Final vs Baseline Farkı:** Başarı farkı nedir ve sizce bu fark nereden kaynaklanıyor?
7. **Business Uyumu:** Bu model iş gereksinimlerine uygun mu? (Örn: Eşik değerini (threshold) değiştirmeli miyim?)
8. **Canlıya Alma ve Monitoring:** Modeli canlıya alsaydık hangi metriklerle izlerdik? Data Drift (veri kayması) olursa nasıl fark ederdik? Bunu düşünmenizi istiyoruz.

6. "Olsa Güzel Olur" (Bonuslar)

Enes Fehmi Manan:

- **Git Geçmişi:** "Upload files" yapmak yerine adım adım commit atılmış düzenli bir git geçmişi.
- **Monitoring Sistemi:** Tahmin sonuçlarının loglandığı (belki SQLite veya basit bir yapı) ve izlendiği bir ekran.
- **Business Kurgusu:** Kendinizi bir şirkette çalışan veri bilimci gibi hayal edip bir sistem tasarımı anlatmak (Örn: Önce banka kurallarına takılıyor, sonra skoringe gidiyor, sonra limit belirleniyor).
- **Üst Yönetim Sunumu:** Teknik olmayan birine (yöneticiye) projenin maliyetini, başarısını ve faydasını anlatan 5-10 slaytlık bir sunum.
- **YouTube Videosu:** Projeyi anlattığınız kısa bir video.
- **Medium Yazısı:** Teknik süreci anlattığınız bir blog yazısı.

7. Soru - Cevap (Q&A)

Soru: Nereden dataset bulabilirim?

Enes & Özge: Kaggle'da arama kısmına sektör yazarak (örn: "Churn", "Credit Risk") "Competitions" sekmesinden

bakabilirsiniz. "Playground" yazanlardan kaçının. Gerçek ödüllü veya geçmiş yarışmalar daha iyidir. Ayrıca Hugging Face Spaces kısmında yapılmış örnek projelere bakarak (örn: "Fraud Detection") ilham alabilirsiniz.

Soru: Kendimiz veri toplayabilir miyiz?
Özge: Evet, mesela e-ticaret verisi bulup kendiniz bir target (hedef değişken) yaratarak (örn: "Gelecek ay ne kadar alacak?") bir veri seti oluşturabilirsiniz. Bu işinizi zorlaştırır ama mülakatlarda anlatırken çok hoşlarına gider.

Soru: Sinyal verisi (Signal Processing) kullanabilir miyiz?
Enes & Özge: Eğitimde tabular (tablo) veri işlediğimiz için sinyal veya görüntü işleme (CNN vb.) konularına girmemenizi öneririz. Tabular veriye sadık kalın.

Soru: Yapay Sinir Ağları (Deep Learning) kullanabilir miyiz?
Enes & Özge: Kullanabilirsiniz, yasak değil. Ancak önce temel Machine Learning algoritmalarıyla (Random Forest, XGBoost vb.) bir baseline kurun. Sonra üzerine deep learning deneyebilirsiniz.

Soru: Sentetik veri ile Monte Carlo analizi yapsak?
Enes: Sentetik datayı önermiyoruz. Çünkü sentetik datada yaptığınız Feature Engineering'in karşılığını alamıyorsunuz. Gerçek hayatın gürültüsünü (noise) içermiyor.

Soru: Kodun görünümü nasıl olmalı?
Enes: Lütfen kodun her yerine emoji, gereksiz HTML süslemeleri veya LLM'den (ChatGPT vb.) kopyalandığı belli olan uzun ve gereksiz yorum satırları **koymayın**. Notebook'unuz "keşkül" gibi uzamasın. Kodunuz **Clean (Temiz)** olsun. Analitik yorumlarınızı aralara yazın, kodun kendisine değil. Amacımız "süslü" notebook değil, "işlevsel ve okunabilir" kod.

Soru: Proje değerlendirmesinde sıralama var mı?
Özge: Hayır, bir yarışma veya sıralama yok. Amacımız bu aşamaları doğru şekilde tamamlamanız ve öğrenmeniz.

Soru: Son teslim tarihi ne zaman?
Enes: 9 Aralık. Yaklaşık 3 haftanız var.

8. Kapanış ve İletişim

Enes Fehmi Manan:
İletişim için WhatsApp grubumuz var. Aklınıza takılan soruları, "Şunu ekledim nasıl olmuş?" gibi paylaşımları oradan yapabilirsiniz. Birbirinizin rakibi değilsiniz, repolarınız public olabilir, yardımlaşın.

Kaynaklar:

- Enes Manan - Credit Risk Model Reposu (Örnek yapı)
- Made With ML (Proje yapısı ve testler için)
- ML Engineering Book (Andriy Burkov) (Mülakatlar ve genel kültür için)
- Data Science Awesome Repo

Hepinize iyi akşamlar, güzel projeler bekliyoruz!