İSTANBUL TOPKAPI ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

FET445 – Veri Madenciliği Final Proje Raporu

Upwork İş Verilerine Dayalı Saatlik Ücret Tahmini

Gurup:Daltons Öğrenciler:

• Ayham Assad – 22040101099

• Abdulkerim Albustani – 22040101100

• Osama Alkheder – 22040101117

• Asil Elnasir – 22040101169

Github: https://github.com/Daltonos-Daltonlar/Upwork-Jobs.git

YouTube Sunum Videosu: https://youtu.be/zS6abppcqds?si=1r-xRpBqv8dYRiUh

1. Özet

Bu rapor, Upwork platformunun 2024 Şubat-Mart dönemindeki iş verilerini kullanarak saatlik oranları tahmin eden makine öğrenmesi modellerinin karşılaştırmalı analizini sunmaktadır. Dört öğrenci tarafından geliştirilen beş farklı kategori altında on bir makine öğrenmesi modeli değerlendirilmiştir. Veri seti 140,936 saatlik ilanı içermektedir ve 1,007-1,507 öznitelik boyutuyla işlenmiştir.

Anahtar Bulgular:

- En yüksek performans: **SVR Linear** ve **Linear Regression** modelleri ($R^2 = 1.0000$)
- En dengeli model: **RandomForestRegressor** ($R^2 = 0.9999$, MAE = 0.0057)
- Derin öğrenme modelleri: MLP ve LSTM modellerinin orta seviye performans gösterdiği gözlenmiştir

2. Giriş

2.1 Problem Tanımı

Freelance platformlarında işçi işvereni ve işçi tarafından belirtilen saatlik ücretler, proje bütçeleri ve diğer faktörler dikkate alınarak belirlenmektedir. Bu çalışmanın amacı, iş ilanı özellikleri kullanarak ortanca saatlik oranını tahmin etmektir. Bu tür tahminler:

- Freelance platformları için dinamik fiyatlandırma stratejileri geliştirmek
- İşçilerin rekabetçi bir oranı belirlemelerine yardımcı olmak
- İşverenlerin bütçe planlaması yapmasını sağlamak

gibi uygulamalara sahiptir.

2.2 Veri Seti Özellikleri

Özellik	Değer
Toplam kayıt	140,936 (sadece saatlik ilanlar)
Öznitelik sayısı	1,007-1,507
Eğitim seti	81,937 örnek
Test seti	20,485 örnek
Hedef değişken	avg_hourly (ortalama saatlik cret)
Veri tarafındaki imbalans	Evet (Ülke dağılımında)

Öznitelik Türleri:

1. **Metin Özellikleri:** TF-IDF ile işlenen iş başlığı (1,000-1,500 öznitelik)
2. **Sayısal Öznitelikler:**
    - Başlık uzunluğu
    - Kelime sayısı
    - Bütçe bilgisi
    - Saatlik oran bilgisi
    - Ülke kodlaması (en sık 10 ülke + Diğer)
3. **İstatistiksel Öznitelikler:** Bütçe doldurma, eksik değer belirteci

3. Metodoloji

3.1 Veri Ön İşleme

Veri ön işleme aşağıdaki adımları içermektedir:

1. **Temizleme:** Boş başlık içeren kayıtlar çıkarıldı
2. **Dönüştürme:** is_hourly değişkeni binary (0/1) formata dönüştürüldü
3. **Filtreleme:** Sadece saatlik ilanlar (is_hourly=1) regresyon analizi için seçildi
4. **Kodlama:** Kategorik değişkenler (ülkeler) sayısal hale getirildi
5. **Vektörleştirme:** Metin özellikleri TF-IDF ile vektörleştirildi
6. **Ölçeklendirme:** PyTorch modelleri için StandardScaler uygulandı

3.2 Model Mimarileri

Dört öğrenci tarafından geliştirilen modeller aşağıda özetlenmiştir:

3.2.1 Osama Alkheder'in Modelleri

- Linear Regression
- XGBRegressor
- MLP (PyTorch: 128→64→1)
- LSTM (PyTorch: LSTM(64)→32→1)

3.2.2 Abdulkerim Albustani'nin Modelleri

- Linear Regression
- Ridge Regression (GridSearchCV ile α=10.0)
- RandomForestRegressor (GridSearchCV ile max_depth=15, min_samples_split=5, n_estimators=100)
- MLP (PyTorch: 128→64→1)

- LSTM (PyTorch: LSTM(64)→32→1)

3.2.3 Asil Elnasir'in Modelleri

- SVR (Linear kernel, C=0.5)
- KNeighborsRegressor (n_neighbors=5, weights='uniform')
- XGBRegressor (hafif ayarlar: max_depth=4, n_estimators=40, learning_rate=0.15)

3.2.4 Ayham Assad'ın Modelleri

(Veriler başarıyla işlenmiş, ancak nihai model sonuçları tam olarak belirtilmemiştir)

3.3 Değerlendirme Metrikleri

Modellerin performansı beş metrik kullanılarak değerlendirilmiştir:

burada gerçek değer, tahmin edilen değer ve örnek sayısıdır.

4. Sonuçlar

4.1 Öğrenci Bazında Performans Tablosu

Öğrenci	Model	R^2	MSE	RMSE	MAE	MAPE
Osama Alkheder	Linear Regression	1.0000	0.0000	0.0001	0.0000	0.0003
	XGBRegressor	0.9422	51.9830	7.2099	0.5586	1.1957
	MLP (128-64-1)	0.9748	22.6535	4.7596	3.3667	17.9444
	LSTM (64-32-1)	0.8758	111.7648	10.5719	4.3712	23.1359
Abdulkerim Albustani	Linear Regression	1.0000	0.0000	0.0001	0.0000	0.0003
	Ridge Regression	1.0000	0.0009	0.0292	0.0250	0.1652
	RandomForest	0.9999	0.1160	0.3405	0.0057	0.0015
	MLP (128-64-1)	0.9798	18.1819	4.2640	3.0275	19.9848
	LSTM (64-32-1)	0.8954	94.1315	9.7021	4.0765	21.6287
Asil Elnasir	SVR (Linear)	1.0000	0.0092	0.0958	0.0956	0.6410
	KNeighborsRegressor	0.9977	2.1004	1.4493	0.3081	1.7486
	XGBRegressor	0.9278	65.0107	8.0629	0.7226	2.7612

Table 1: Tüm Modellerin Regresyon Performans Metrik Karşılaştırması

4.2 En İyi Modeller

Tablo 2 de gösterildiği üzere, en iyi R^2 skorlarına sahip üç model:

1. **Linear Regression & Ridge Regression & SVR Linear** ($R^2 = 1.0000$)
    - Mükemmel uyum sağlayan bu modeller veri seti içinde çok güçlü doğrusal ilişkiler içerdiğini gösterir

- Gerçek dünya uygulamalarında bu türü yüksek uyumlar overfitting olasılığı taşır

2. **RandomForestRegressor** (R^2 = 0.9999, MAE = 0.0057)
    - Düşük MAE değeri ile çok düşük mutlak hata gösterir
    - Ensemble yöntemi sayesinde istikrarlı tahminler sunar
    - Overfitting riski nispeten düşüktür
3. **KNeighborsRegressor** (R^2 = 0.9977, MAE = 0.3081)
    - Yüksek R^2 değeri ile güçlü tahmin gücü gösterir
    - Dengeli hata metrik gösterir

4.3 Model Kategorisine Göre Analiz

4.3.1 Doğrusal Modeller

Model	R^2	MSE	RMSE	MAE	MAPE
Linear Regression	1.0000	0.0000	0.0001	0.0000	0.0003
Ridge Regression	1.0000	0.0009	0.0292	0.0250	0.1652
SVR (Linear)	1.0000	0.0092	0.0958	0.0956	0.6410
Ortalama	**1.0000**	**0.0034**	**0.0417**	**0.0402**	**0.2688**

Table 2: Doğrusal Modellerin Performansı

Gözlem: Doğrusal modeller mükemmel performans göstermiştir. Bu, veri setinde güçlü doğrusal ilişkiler bulunması veya veri seti yapısının bu türü modellere uygun olması anlamına gelmektedir.

4.3.2 Ağaç Tabanlı Modeller

Model	R^2	MSE	RMSE	MAE	MAPE
RandomForestRegressor	0.9999	0.1160	0.3405	0.0057	0.0015
XGBRegressor (Osama)	0.9422	51.9830	7.2099	0.5586	1.1957
XGBRegressor (Asil)	0.9278	65.0107	8.0629	0.7226	2.7612
Ortalama	**0.9566**	**39.0366**	**5.2044**	**0.4290**	**1.3195**

Table 3: Ağaç Tabanlı Modellerin Performansı

Gözlem: RandomForest mükemmel performans gösterirken, XGBRegressor modelleri daha düşük performans sergilemiştir. Bu, hyperparameter ayarlanması (GridSearchCV vs hafif ayarlar) ve eğitim veri boyutu seçiminin önem taşıdığını gösterir.

4.3.3 Derin Öğrenme Modelleri (PyTorch)

Model	R^2	MSE	RMSE	MAE	MAPE

MLP Osama (128-64-1)	0.9748	22.6535	4.7596	3.3667	17.9444
LSTM Osama (64-32-1)	0.8758	111.7648	10.5719	4.3712	23.1359
MLP Abdulkerim (128-64-1)	0.9798	18.1819	4.2640	3.0275	19.9848
LSTM Abdulkerim (64-32-1)	0.8954	94.1315	9.7021	4.0765	21.6287
Ortalama	**0.9315**	**61.6829**	**7.3244**	**3.9105**	**20.6779**

Table 4: Derin Öğrenme Modellerin Performansı

Gözlem:

- MLP modelleri LSTM modellerinden üstün performans göstermiştir (R^2 MLP: 0.97+ vs LSTM: 0.87-0.89)
- LSTM modellerinin zayıf performansı, veri setinin zamansal olmayan yapısı nedeniyle LSTM'nin avantajlarını kullanamadığını gösterir
- Derin öğrenme modelleri geleneksel modellere kıyasla daha düşük R^2 değerleri göstermiştir

4.3.4 Komşu Tabanlı Modeller (KNN)

Model	**R^2**	**MSE**	**RMSE**	**MAE**	**MAPE**
KNeighborsRegressor	0.9977	2.1004	1.4493	0.3081	1.7486

Table 5: KNN Modeli Performansı

Gözlem: KNN modeli çok güçlü performans göstermiş, $R^2 = 0.9977$ ve düşük MAE değerleri ile standart regresyon uygulamaları için uygun bulunmuştur.

4.4 Modellerin Genel Sıralaması

En iyi 5 model şu şekilde sıralanmaktadır:

Sıra	Model	Öğrenci	R^2	MAE
1	Linear Regression	Osama/Abdulkerim	1.0000	0.0000
2	Ridge Regression	Abdulkerim	1.0000	0.0250
3	SVR (Linear)	Asil	1.0000	0.0956
4	RandomForestRegressor	Abdulkerim	0.9999	0.0057
5	KNeighborsRegressor	Asil	0.9977	0.3081

5. En İyi Model Analizi: RandomForestRegressor

RandomForestRegressor modeli (Abdulkerim Albustani'nin çalışması) genel olarak en dengeli ve güvenilir model olarak belirlenmiştir.

5.1 Neden RandomForestRegressor?

1. **Mükemmel R² Değeri:** R² = 0.9999 (doğrusal modellere çok yakın)
2. **En Düşük MAE:** MAE = 0.0057 (en düşük ortalama mutlak hata)
3. **Düşük Overfitting Riski:** Ensemble yöntemi sayesinde genelleme yeteneği iyi
4. **Hyperparameter Ayarlanması:** GridSearchCV ile optimal parametreler seçilmiş
5. **Gerçekçi Performans:** Doğrusal modellerin 1.0000 R² değerleri overfitting gösterirken, RF daha gerçekçi bir uyum sağlamıştır

5.2 Hiperparametreler

RandomForestRegressor optimal ayarları:

- **n_estimators:** 100
- **max_depth:** 15
- **min_samples_split:** 5
- **random_state:** 42

Bu parametreler GridSearchCV kullanılarak otomatik olarak seçilmiştir.

5.3 Model İstatistikleri

Metrik	Değer
R² Skoru	0.9999
MSE	0.1160
RMSE	0.3405
MAE	0.0057
MAPE	0.0015

Table 6: RandomForestRegressor Detaylı Performans

6. Tartışma

6.1 Temel Bulgular

1. **Doğrusal Modellerin Üstünlüğü:** Veri seti doğrusal regresyon modelleri tarafından neredeyse kusursuz bir şekilde açıklanabilmektedir. Bu, özniteliklerin hedef değişkenle güçlü doğrusal ilişkiye sahip olduğunu gösterir.

2. **Derin Öğrenme Zorlukları:** LSTM modelleri, zamansal özelliği olmayan veri seti üzerinde etkisiz olmuştur. MLP modelleri daha iyi performans göstermesine rağmen, geleneksel yöntemler kadar iyi sonuç almamıştır.

3. **Ensemble Yöntemlerinin Başarısı:** RandomForest, çok iyi hiperparameter ayarlaması ile mükemmel sonuçlara ulaşmıştır.

4. **Hyperparameter Ayarlanması:** GridSearchCV kullananların (Abdulkerim) sonuçları, sabit parametreler kullananlardan (Asil) daha iyi olmuştur.

6.2 Öğrenci Performansları

- **Abdulkerim Albustani:** 5 model geliştirmiş, çeşitli yöntemler denemiş, GridSearchCV kullanmıştır
- **Osama Alkheder:** 4 model geliştirmiş, PyTorch ile derin öğrenme uygulamıştır
- **Asil Elnasir:** 3 model geliştirmiş, hafif ayarlarla hızlı eğitim tercih etmiştir
- **Ayham Assad:** Veri hazırlanması tamamlanmıştır

6.3 Sınırlamalar ve Gelecek Çalışmalar

1. **Overfitting Şüphesi:** Çok yüksek R^2 değerleri overfitting olasılığını işaret etmektedir
2. **Zamansal Analiz:** Veri setine zamansal boyut eklenerek LSTM etkinliği test edilebilir
3. **Feature Mühendisliği:** Daha gelişmiş öznitelik oluşturma teknikleri uygulanabilir
4. **Çapraz Doğrulama:** K-fold çapraz doğrulama yapılarak modeller daha güvenilir şekilde değerlendirilmelidir

7. Sonuç

Bu çalışmada, Upwork platformunun saatlik iş verilerine dayalı olarak, on bir makine öğrenmesi modeli değerlendirilmiştir. Bulgular göstermiştir ki:

1. **En iyi model: RandomForestRegressor** ($R^2 = 0.9999$, MAE = 0.0057) - Abdulkerim Albustani'nin çalışması
2. **Yüksek performans:** Doğrusal modeller de mükemmel sonuçlar vermiştir ($R^2 = 1.0000$)
3. **Moderate performans:** Derin öğrenme modelleri orta seviye sonuçlar sağlamıştır ($R^2 = 0.87$-0.97)

Veri seti yapısı, doğrusal ve ağaç tabanlı yöntemlerin çok etkili olması için uygun bulunmuştur. RandomForestRegressor, mükemmel performans ve düşük overfitting riski nedeniyle prodüksiyon ortamında kullanılması için önerilmektedir.

Kaynakça

[1] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[3] Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

[4] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.