İSTANBUL TOPKAPI ÜNİVERSİTESİ

FET445 Veri Madenciliği / Güz Dönemi

Proje Başlığ : Upwork Serbest Çalışma Piyasası Analizi ve Yetenek Talebi Tahmini

DALTONS ÜYELERİ

ADI	SOYADI	NO	e-mail
OSAMA	ALKHEDER	22040101117	osamaalkheder@stu.topkapi.edu.tr
AYHAM	ASSAD	22040101099	ayhamasasd@stu.topkapi.edu.tr
ABDULKERIM	ALBUSTANI	22040101100	abdalkarreemalbustani@stu.topkapi.edu.tr
ASIL	ELNASIR	22040101196	asilelnasir@stu.topkapi.edu.tr

GitHub/Repo link: https://github.com/Daltonos-Daltonlar/Upwork-Jobs.git

1. Problem Tanımı

Bu proje, Upwork üzerinde yayınlanmış 200.000'den fazla serbest iş ilanını analiz ederek veri bilimi, yazılım geliştirme, grafik tasarım ve benzeri alanlarda en çok talep edilen becerileri belirlemeyi amaçlar.
Ayrıca iş ilanlarını kategorilere göre sınıflandırmak (classification) ve bütçe tahmini yapmak (regression) hedeflenmektedir.

2. Veri Seti Tanımı

Veri seti: all_upwork_jobs_2024-02-07-2024-03-24.csv
Değişkenler: title, description, skills, category2, country, budget.
Toplam kayıt: ~200K

3. Örnek Veri Satırları

1) Python developer for data scraping – Budget: 180 USD2) React + Node full-stack developer – Budget: 450 USD

3) Logo design project – Budget: 75 USD

4) NLP classification ML engineer – Budget: 300 USD

5) Shopify customization expert – Budget: 250 USD

4. Keşifsel Veri Analizi (EDA)

- Eksik veri analizi (NA oranları)
- Dağılım grafiklerinin incelenmesi (histogram, bar charts)- Top 20 ülke, Top 20 kategori,

Top 30 skill- Aykırı değer tespiti (budget)
- Sızıntı risk kontrolü

5. Veri Hazırlama

- Eksik veri imputasyonu
- Text temizleme: lower-case, stopword removal
- TF-IDF vektörizasyon (title + description + skills)
- Kategorik değişken kodlama (One-Hot)
- Sayısal değişken standardizasyonu

6. Özellik Mühendisliği

- description_len, title_len
- Kelime sayısı
- TF-IDF bileşenleri
- PCA ile boyut indirgeme (50, 100, 150 bileşen karşılaştırması)

7. Modelleme Planı

- Baseline: Majority classifier (CLS), Mean predictor (REG)
- Classification Modelleri: Logistic Regression, Random Forest, KNN
- Regression: RandomForestRegressor
- Stratified 80/20 train-test split

8. Hiperparametre Optimizasyonu

GridSearchCV uygulanmıştır:
- Logistic Regression → C parametresi
- Random Forest → n_estimators, max_depth
- Performans ölçütü: macro-F1 (classification), RMSE (regression)

9. Değerlendirme Tasarımı

Classification: Accuracy, Precision, Recall, F1, Confusion Matrix
Regression: RMSE, MAE
CV: 5-fold stratified cross-validation

10. Riskler ve Azaltma

- Veri dengesizliği → SMOTE, class_weight
- Overfitting → regularization
- Aykırı değerler → Winsorization
- Runtime yüksekliği → PCA, daha küçük modeller

11. Kullanılan Araçlar ve Ortam

Python 3.x, pandas, numpy, scikit-learn, matplotlib, seaborn
Notebook dosyaları: 1_data_cleaning.ipynb, 2_eda_visualization.ipynb,
3_model_building.ipynb, 4_hyperparameter_tuning.ipynb

12. Beklenen Sonuçlar

- En çok talep edilen 20 skill'in grafikleri
- Category2 için sınıflandırma sonuç tablosu
- Budget tahmini için RMSE tablosu- PCA sonrası model performans karşılaştırma tablosu
- ROC, PR eğrileri

13. Referanslar

[1] Upwork Job Dataset (Kaggle)
[2] Scikit-Learn Documentation[3] TF-IDF Research Papers
[4] Logistic Regression & Random Forest Academic Sources

14. Proje Yönetimi & Zaman Çizelgesi

1. Hafta: Veri seti seçimi ve proje konusunun belirlenmesi

2. Hafta: Veri ön işleme + EDA başlangıcı

3. Hafta: EDA tamamlanması + Baseline model geliştirme

4. Hafta: İleri modelleme (LR, RF, KNN) + PCA

5. Hafta: Hyperparameter tuning

6. Hafta: Performans analizi + hata analizi

7. Hafta: Nihai rapor ve sunum hazırlığı

15. Roller ve Sorumluluklar

Abdulkerim: Veri temizleme ve ön işleme

Ayham: EDA ve görselleştirme

Asil: Baseline modeller ve temel model karşılaştırması

Osama: PCA, tuning ve nihai model optimizasyonu

16. Proje Çıktıları

- Final proje raporu

- 5 adet Jupyter Notebook

- Veri sözlüğü

- Sunum dosyası (Word)

- GitHub deposu

17. İlgili Çalışmalar (Literatür İncelemesi)

Literatürde Upwork ve freelancer piyasası üzerine yapılan çalışmalar genellikle beceri talebi, ücret tahmini ve iş kategorisi sınıflandırmasına odaklanmaktadır.
1) "Freelance Market Analysis using Machine Learning" – Kaggle Blog
2) "Skill Demand Forecasting in Online Job Markets" – IEEE
3) "Text Classification for Job Category Prediction" – Medium ML Article
Bu proje, önceki çalışmalardan farklı olarak hem kategori tahmini hem de bütçe regresyonu içeren çift görevli bir yaklaşım sunmaktadır.

18. Veri Yönetimi, Etik ve Fairness

Veri Kaynağı: Upwork Dataset (Kaggle)
Lisans: Kamuya açık veri kaynağı
Etik: Kişisel veriler temizlenmiştir. Hassas veri yoktur.
Fairness: Ülke değişkeninde önyargı riskleri değerlendirilmiştir.Erişim Planı: CSV dosyası /data klasöründe saklanmakta ve notebooklar tarafından otomatik yüklenmektedir.

19. Veri Sözlüğü (Data Dictionary)

title: İş ilanı başlığı (string)
description: İş tanımı (string)
skills: Gerekli beceriler (string list)
category2: Upwork kategorisi (string)
country: İlan sahibinin ülkesi (string)
budget: Bütçe (numerik USD)