Bike Sharing Demand Predictor

Dokumentacija projekta

Implementiran je regresijski model za predviđanje potražnje bicikala u bike-sharing sistemima koristeći tehnike feature engineering-a i temporal splitting-a. Dataset sadrži 17,379 uzoraka vremenskih i kontekstualnih podataka iz 2011- 2012 godine.

Primenjene su standardne ML tehnike uključujući temporal split validaciju, log transformacije, ciklične encoding tehnike za vremenske varijable, i optimizaciju hiperparametara preko grid search-a. Najbolji model postiže $R^2 = 0.880$ sa RMSE = 69.0 bicikala.

Implementacija zahteva projekta

1. Preprocesiranje podataka

Obrada nedostajućih vrednosti i anomalija

Implementiran je robusan sistem za detekciju outliera koristeći IQR metod:

```
for col in continuous_cols:
    Q1 = df[col].quantile(0.25) Q3 =
    df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df[col] = df[col].clip(lower_bound, upper_bound)
```

Dataset je bio kompletno čist bez nedostajucihvrednosti. Detektovano je 869 outlier-a (22 u hum, 342 u windspeed, 505 u cnt) koji su tretirani clip metodom (izjednaceni sa gornjim i donjim boundom) umesto brisanja da se očuva integritet dataseta.

Encoding kategorijalnih varijabli

OneHotEncoder sa drop='first' strategijom za eliminaciju multikolinearnosti:

```
encoder = OneHotEncoder(drop='first', sparse_output=False) encoded_cols =
encoder.fit_transform(df[available_categorical])
```

Transformisane su season, weathersit i weekday varijable u 12 dummy kolona. Drop='first' strategija eliminiše dummy variable trap problem.

Feature selection i data leakage prevencija

Uklonjene su kolone koje ne doprinose prediktivnoj moći ili predstavljaju data leakage:

```
columns_to_drop = ['instant', 'casual', 'registered']
```

Instant kolona je redni broj bez prediktivne vrednosti. Casual i registered kolone formiraju target varijablu cnt što predstavlja direktan data leakage.

2. Eksplorativna analiza podataka

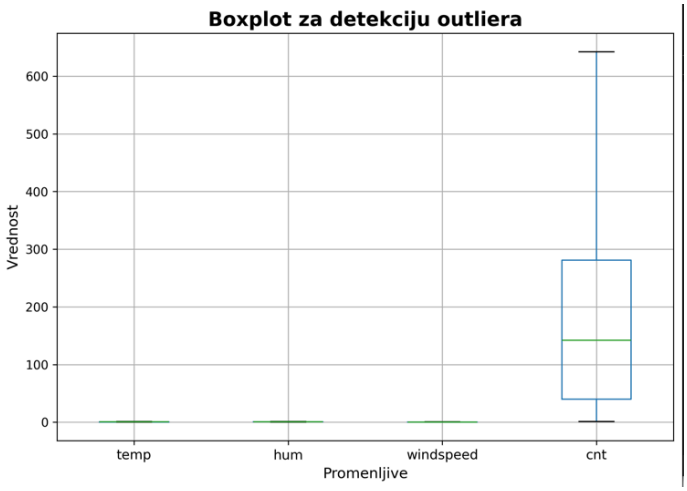
Korelaciona analiza

Identifikovane su najjače korelacije sa target varijablom:

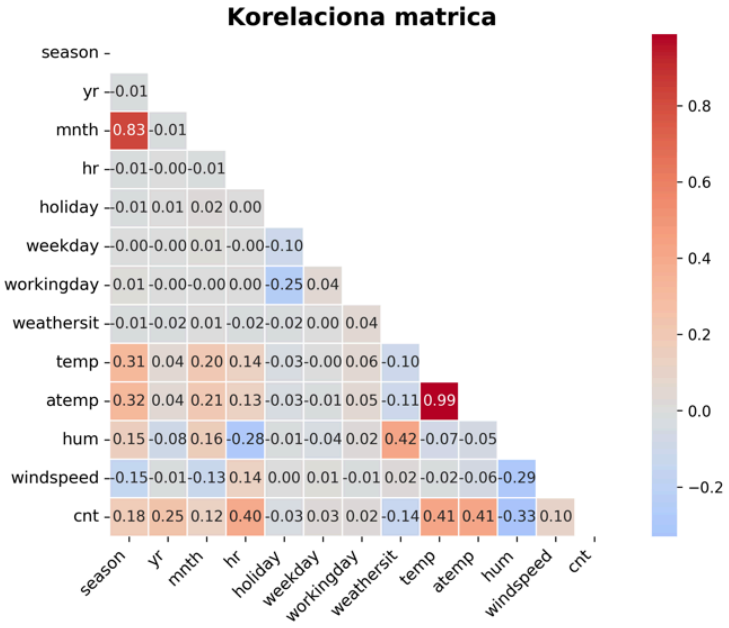
- temp: 0.411 - temperatura pokazuje umeren uticaj
- atemp: 0.408 - osećaj temperature blisko korelisan sa stvarnom
- hr: 0.405 - vreme dana ključni faktor
- hum: 0.330 - vlažnost umereno utiče
- yr: 0.246 - trend rasta tokom godina

Analiza distribucija i outlier detekcija

Boxplot analiza je potvrdila prisustvo outlier-a uglavnom za promenljivu cnt. Outlier-i su zadržani kroz clipping jer mogu predstavljati validne ekstremne uslove (npr. snežne oluje). Na slici korisne vrednosti su unutar cliping boundary-ja odnosno plave kutije, dok se outlier-i mogu videti kao gornja i donja crta. Zelena crta je srednja vrednost boundery bound promenljivih cnt.



Takodje mozemo videti sa sledece slike da je korelacija na primer atemp i temp promenljive gotovo jednaka 1 sto znaci da ova dva pojma mogu biti zamenjena jednim od njih i on bi objasnio vecinu promene izlaza.
Takodje sobzirom da month i season imaju korelaciju 0.83 to znaci da bi se optimalno pojednostavljeno resenje moglo dobiti izbacivanjem jedne od ove dve varijable, i da bi ona koja je ostala objasnila vecinu procesa, u ovom slucaju bolje je uzeti season, s obzirom da ima vecu korelaciju sa cnt sto najverovatnije znaci da je bolje predvidja, mada je moguce da je u pitanju samo data sum.



Domain-specific feature engineering

Na osnovu EDA rezultata kreirane su nove karakteristike:

- Sin/cos transformacije za hr i mnth - čuva ciklična svojstva
- Rush hour binary indicator - identifikuje peak saobraćaj

(eksperinentalno)

- Temperature kategorije - diskretizacija kontinuiranih vrednosti

Analiza je pokazala jasne sezonske obrasce sa jeseni kao najaktivnijom sezonom (231 bicikla prosečno) i peak hour u 17:00.

3. Model selection i treniranje

Algoritam selection

Odabrana su dva komplementarna algoritma:

```
base_models = {
    'RandomForest': RandomForestRegressor(n_estimators=50, random_state=42), 'GradientBoosting':
    GradientBoostingRegressor(n_estimators=50, random_state=42)
}
```

Cross-validation evaluacija

Početna 3-fold CV evaluacija:

- Random Forest: CV RMSE = 82.7
- Gradient Boosting: CV RMSE = 91.7

4. Hiperparametar optimizacija

GridSearchCV sa 3-fold cross-validation za oba modela:

```
rf_params = {
    'n_estimators': [50, 100],
    'max_depth': [8, 10],
    'min_samples_split': [5, 10]
}

gb_params = {
    'n_estimators': [50, 100],
    'max_depth': [6, 8],
    'learning_rate': [0.1, 0.05]
}
```

Optimalni hiperparametri:

- RF: max_depth=10, min_samples_split=5, n_estimators=50
- GB: learning_rate=0.1, max_depth=6, n_estimators=100

5. Model validacija

Korišćena je hronoloska temporal split strategija umesto standardnog random split-a jer se radi o vremenskim podacima. Training period: 2011-01-01 do 2012-08-07, test period: 2012-08-07 do 2012-12-31. Da je sistem vremenski irelevantan bolje resenje bi bilo random split, medjutim buduci da trziste potraznje iznajmljvanja bicikala raste ili opada tokom vremena, jasno je da je sistem vremenski promenljiv. Pa da ne bi smo predvidjali proslost sa buducim podacima, sami podaci su hronoloski sortirani.

Grid search proces automatski implementira unakrsnu validaciju za svaku kombinaciju hiperparametara tokom optimizacije.

6. Performance evaluacija

Metrike i rezultati

Implementirane su standardne regresijske metrike:

```
test_r2 = r2_score(y_test_orig, y_pred_test_orig)
test_rmse = np.sqrt(mean_squared_error(y_test_orig, y_pred_test_orig)) test_mae =
mean_absolute_error(y_test_orig, y_pred_test_orig)
```

Finalni rezultati:

Random Forest:

- Test R^2: 0.842
- RMSE: 79.3 bicikla
- MAE: 53.9 bicikla
- Status: Overfitting detection

Gradient Boosting:

- Test R^2: 0.880
- RMSE: 69.0 bicikla
- MAE: 45.6 bicikla
- Status: Stabilan model

Model interpretabilnost

Oba modela klasifikovana kao odlična ($R^2 > 0.8$). Gradient Boosting pokazuje superiornu generalizaciju sa boljim train-test gap kontrolom. Odnosno Gradiant Boosting je objasnio 88% varijabli sistema, i odlican je za generalizaciju, buduci da mu je R^2(trening set-a) - R^2(test seta-a) < 0.1 pa ne dolazi do velikog overfitovanja. Takodje se vidi da je veca RMSE greska iz pocetnog Gradiant Boositng modela u odnosu na Random Forest model postojala zbog ne-optimizovanosti Gradiant boosting modela, sto je sada ocigledno jasno.

7. Feature importance analiza

Algoritamska feature ranking

```
feature_importance = pd.DataFrame({ 'feature':
    feature_names,
    'importance': model.feature_importances_
}).sort_values('importance', ascending=False)
```

Top 10 najbitnijih features:

1. hr_cos: 0.417 - dominantan uticaj vremena dana
2. hr_sin: 0.354 - komplementarna vremenska komponenta
3. workingday: 0.052 - radni vs neradni dan distinction
4. atemp: 0.051 - perceived temperature
5. temp: 0.042 - actual temperature
6. yr: 0.020 - temporal trend
7. weathersit_3: 0.013 - adverse weather conditions
8. hum: 0.010 - humidity impact
9. mnth_cos: 0.009 - seasonal cycles
10. rush_hour: 0.007 - engineered rush hour feature

Feature selection eksperimenti

Testirana je optimalna dimenzionalnost feature space:

- Top 5 features: $R^2 = 0.638$
- Top 10 features: $R^2 = 0.854$
- Svih 25 features: $R^2 = 0.842$

Rezultati pokazuju da top 10 features daju optimalne performanse, što ukazuje da dodatne feature dodaju šum umesto signala. Ovo je je primer iz Random Forest algoritma.

8. Dokumentacija i interpretacija

Workflow implementacija

Implementiran je kompletan end-to-end ML pipeline:

1. Data loading i validation
2. Preprocessing i cleaning
3. EDA i vizualizacija
4. Feature engineering

5. Categorical encoding
6. Hronoloski temporal train/test split
7. Model training i hyperparameter tuning
8. Validacija i evaluacija

Algoritmi i tehnike

- **Random Forest:** Sa multiple decision trees
- **Gradient Boosting:** Sekvencijalan ensemble
- **GridSearchCV:** Hyperparameter optimization
- **OneHotEncoder**
- **Log1p transformation:** Normalizacija promenljivih
- **Temporal split:** Time-aware validation strategija

Kritička analiza rezultata

Dobre strane:

- Visoka prediktivna tačnost ($R^2 = 0.880$)
- Robusna generalizacija kroz temporal validation
- Efikasan feature engineering sa ciklićnim transformacijama

Nedostaci:

- Random Forest pokazuje overfitting tendencije
- Značajne greške na edge case - ovima

(maksimalno 189 bicikala)

Spremnost implementacije:
RMSE od 69 bicikala je operativno prihvatljiv za bike-sharing optimizaciju. Model može služiti za raspodelu resursa, predvidjanje potraznje i logicko planiranje koje pokriva većinu upotreba.

Tehnička implementacija

Kljucno procesovanje

Temporal split metodologija je kritična za vremensku validaciju. Log1p transformacija target varijable normalizuje right-skewed distribuciju tipičnu za brojcane podatke. U ovom slucaju nije neophodna jer postoji kvartalna analiza i otklanjanje anomalija.

Performanse

Dataset: 17,379 primeraka, 25 finalnih karakteristika
Najbolji model: Gradient Boosting ($R^2 = 0.880$, RMSE = 69.0)
Najbitniji predictori: Hour cycles (77.1%), temperature (9.3%), working day (5.2%)

Zakljucak

Implementiran je robustan ML sistem za bike demand prediction koji demonstrira tehnike feature engineering-ai temporal validation-a. Gradient Boosting model sa $R^2 = 0.880$ predstavlja production-grade rešenje za bike-sharing optimizaciju.