ZNEUS – PROJEKT1 – ADRIAN URBANEK

Predikcia median_house_value

HOUSES

Dátový zdroj: dataset s 20 640 riadkami a 9 číselnými stĺpcami; bez chýbajúcich hodnôt a bez duplicitných riadkov. Cieľ **median_house_value**, 965 hodnôt je cenzorovaných na maxime 500 001.

1) Zhrnutie dát a motivácia

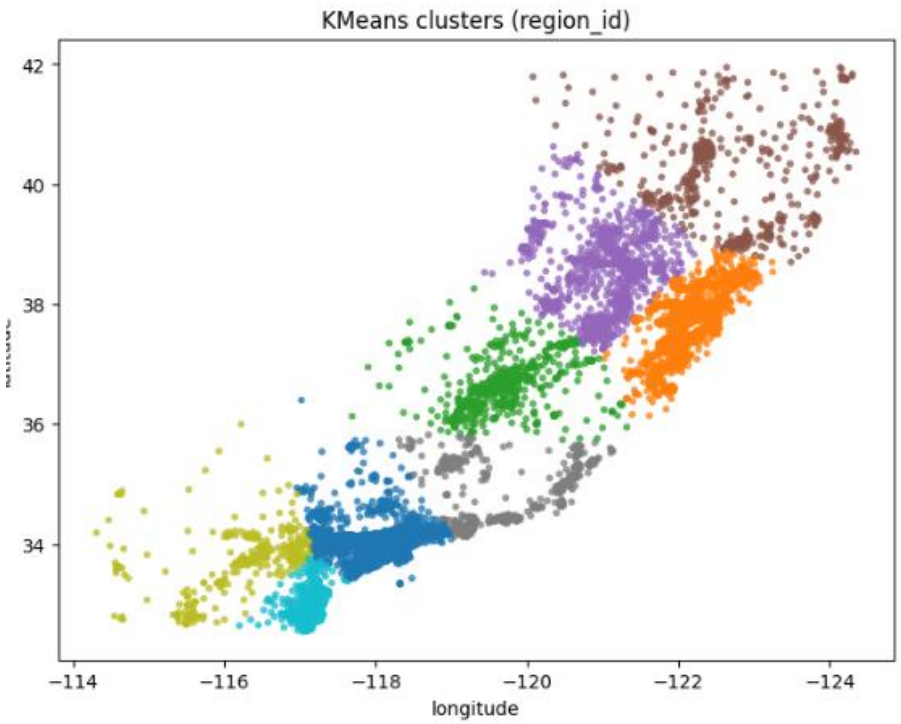
- Pracujem s čisto numerickými premennými: príjem, vek bývania, izby/ložnice/populácia/gospodárstva, zemepisná šírka/dĺžka. Základné štatistiky potvrdzujú veľký rozptyl cieľa. Najsilnejší lineárny signál voči cieľu má median_income (~0.69). Viaceré „count“ premenné majú vysokú šikmosť (skew) a ťažké chvosty

```
Top features by absolute correlation with target:
median_income         0.688075
latitude              0.144160
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.050594
longitude             0.045967
population            0.024650
Name: median_house_value, dtype: float64
```

```
skew = num.skew().sort_values(ascending=False)
display(skew)
```

```
population            4.935858
total_rooms           4.147343
total_bedrooms        3.453073
households            3.410438
median_income         1.646657
median_house_value    0.977763
latitude              0.465953
housing_median_age    0.060331
longitude            -0.297801
dtype: float64
```

- V EDA som si všimol výrazny priestorový vzor (KMeans k≈8 nad lat/lon vytvára regióny s rôznymi mediánmi), preto som do neskorších experimentov pridal geo-features



2) Základný pipeline (spoločný pre všetky experimenty)

1. **Split (80/10/10)** s kvantilovou stratifikáciou na cieľ

```
Train: (16512, 8),  Val: (2064, 8),  Test: (2064, 8)
```

2. **Výber a príprava vstupov:**
    - 8 numerických stĺpcov (viď vyššie).
    - Vybrané stĺpce so |skew| ≥ 1 → log1p (počítané **len na TRAIN**).
    - StandardScaler (fit **len na TRAIN**).

```
Target: median_house_value
Počet numerických features: 8
Features: ['median_income', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households', 'latitude', 'longitude']
```

```python
def suggest_log_cols(frame: pd.DataFrame, cols: list[str], skew_threshold: float = 1.0) -> list[str]:
    skew_vals = frame[cols].skew(numeric_only=True)
    return skew_vals[skew_vals.abs() >= skew_threshold].index.tolist()

def make_preprocess(feature_cols: list[str], log_cols: list[str] | None = None):
    if log_cols is None:
        log_cols = []
    other_cols = [c for c in feature_cols if c not in log_cols]

    col_tf = ColumnTransformer(
        transformers=[
            ("log", FunctionTransformer(np.log1p, validate=False), log_cols),
            ("num", "passthrough", other_cols),
        ],
        remainder="drop"
    )
    preprocess = Pipeline([
        ("cols", col_tf),
        ("scaler", StandardScaler()),
    ])
    return preprocess
```

```python
LOG_COLS_SUGGESTED = suggest_log_cols(df, FEATURES_NUM, skew_threshold=1.0)
print("Návrh log1p stĺpcov (na celej vzorke):", LOG_COLS_SUGGESTED)
```

```
Návrh log1p stĺpcov (na celej vzorke): ['median_income', 'total_rooms', 'total_bedrooms', 'population', 'households']
```

3. **Model:** MLP pre tabuľkové dáta (PyTorch), s možnosťou:
    - **BatchNorm**, **Dropout**, **bottleneck**, **reziduálne bloky**
4. **Tréning:**
    - Optimizer **Adam** (používal som aj RMSprop/SGD v searchoch)
    - **ReduceLROnPlateau** na validačný loss
    - **Early stopping** podľa validačného lossu
    - **Grad clipping** a **weight decay** (pri niektorých nastaveniach)
    - **AMP** na GPU (rýchlosť/nízka pamäť)

Anti-overfit opatrenia: train-only fit (log1p/scaler), validácia oddelená od testu, early stopping, dropout, weight decay.
Anti-underfit: adekvátna kapacita MLP, batch-norm, jemné dolaďovanie LR a batch size.

3) Prehľad experimentov

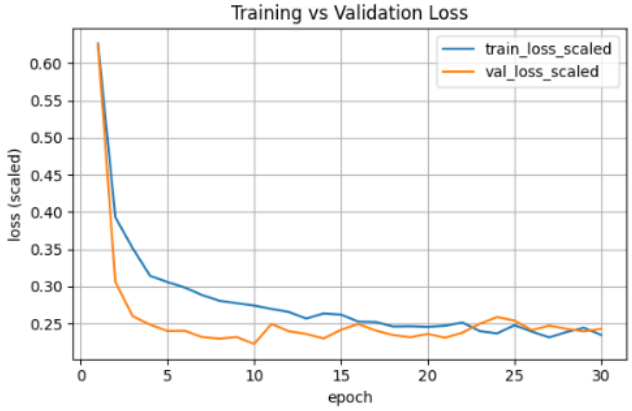
EXPERIMENT 1 — All + (log1p na šikmé)

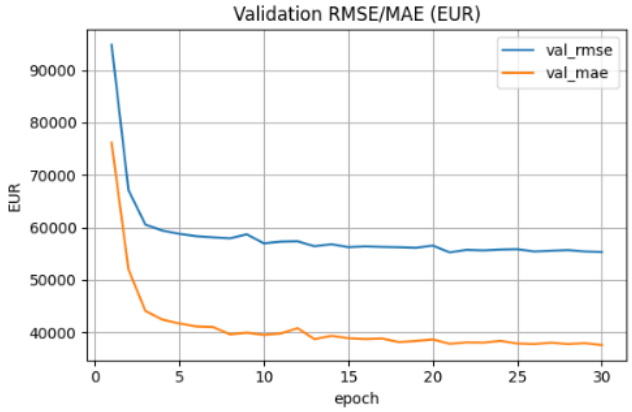
Čo som spravil

- Vstupy len 8 numerických stĺpcov (bez akýchkoľvek geografických konštruktov).
- log1p na stĺpcoch s |skew| ≥ 1 (typicky: median_income, total_rooms, total_bedrooms, population, households).
- Štandardizácia, MLP [256,128,64], BN + Dropout ~0.10, Adam 1e-3, early stopping.

Ako to dopadlo (približne)

- Výsledky: **Train** ~54k, **Val** ~57k / ~40k MAE / R^2 ~0.75–0.76, **Test** ~57k / ~40k MAE / R^2 ~0.75
- Test ≈ Val → stabilné, ale bez extra presnosti



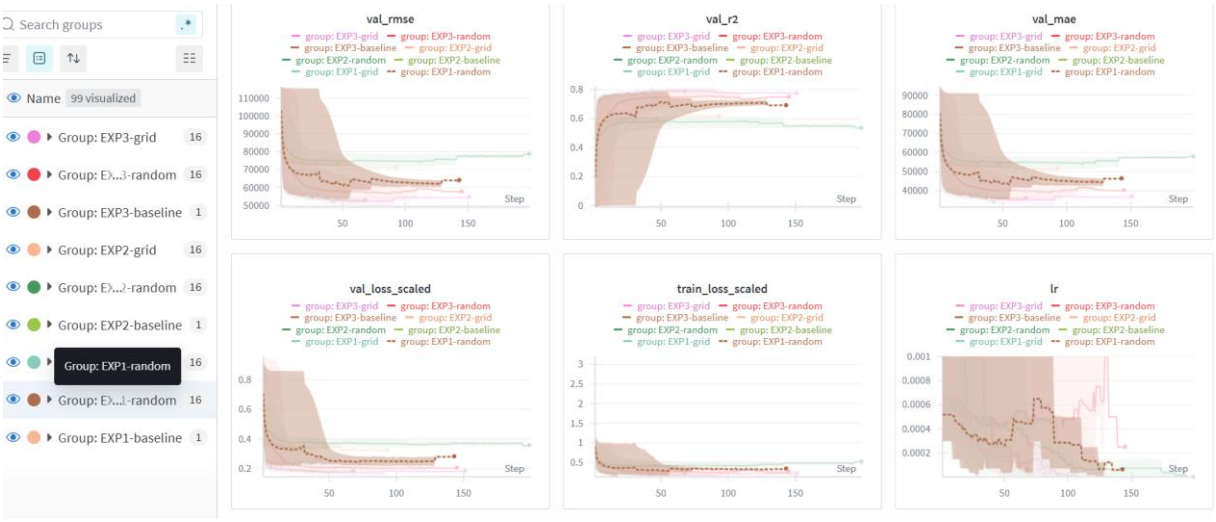



```
]: print(pd.DataFrame(res_baseline).round(4))
```

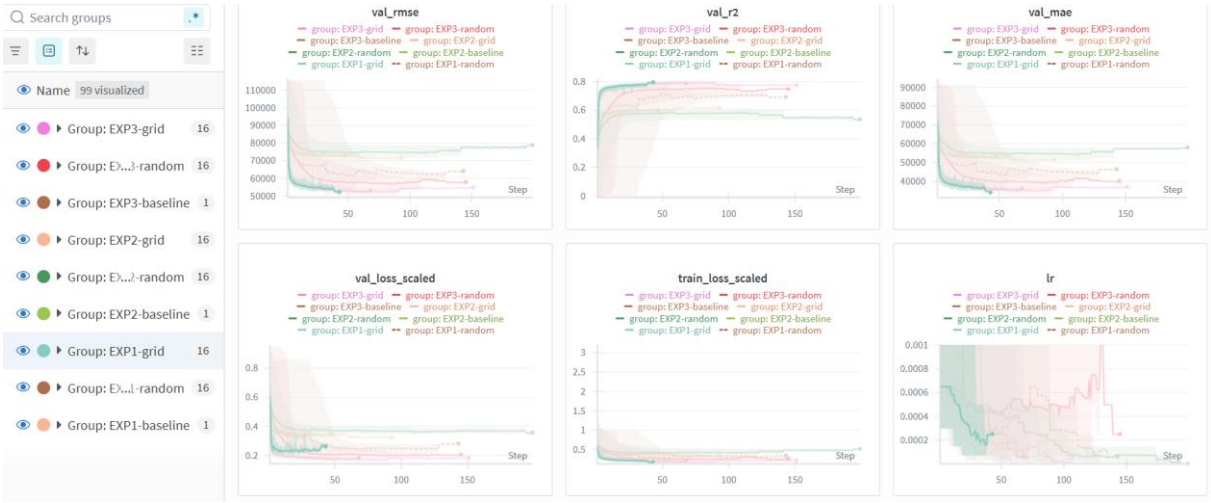
```
          train         val        test
rmse  54497.3398  57005.7695  57039.6641
mae   38448.7500  39567.6328  39544.8516
r2        0.7771      0.7575      0.7529
```

WANDB:

RANDOM:



GRID:



EXPERIMENT 2 — Výber najlepších vstupov podľa |corr|

Čo som spravil

- Na **TRAIN** som vypočítal korelácie všetkých kandidátov s cieľom a vybral **Top-4** stĺpce podľa |corr|
- Na tieto Top-4 som opäť aplikoval log1p podľa šikmosti a následne škálovanie
- Model MLP ako v EXP1; skúšal som aj **random/grid search** nad architektúrou a tréningom

```
TOP4_E2 = corr_tbl.abs().sort_values(ascending=False).head(4).index.tolist()
print("EXP2 Top-4 (TRAIN |corr|):", TOP4_E2)
```

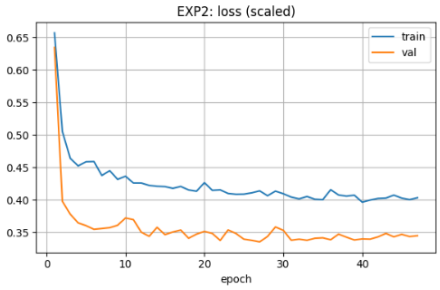
```
EXP2 Top-4 (TRAIN |corr|): ['median_income', 'latitude', 'total_rooms', 'housing_median_age']
```

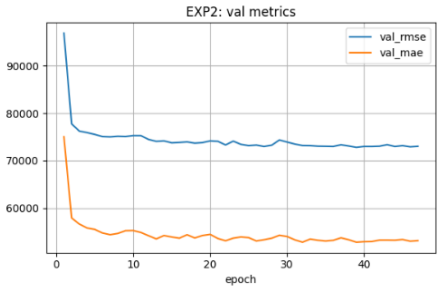
```
LOG_E2 = X_train2[TOP4_E2].skew(numeric_only=True).abs().pipe(lambda s: s[s >= 1.0].index.tolist())
print("EXP2 log_cols (TRAIN):", LOG_E2)
```

```
EXP2 log_cols (TRAIN): ['median_income', 'total_rooms']
```

Ako to dopadlo (približne)

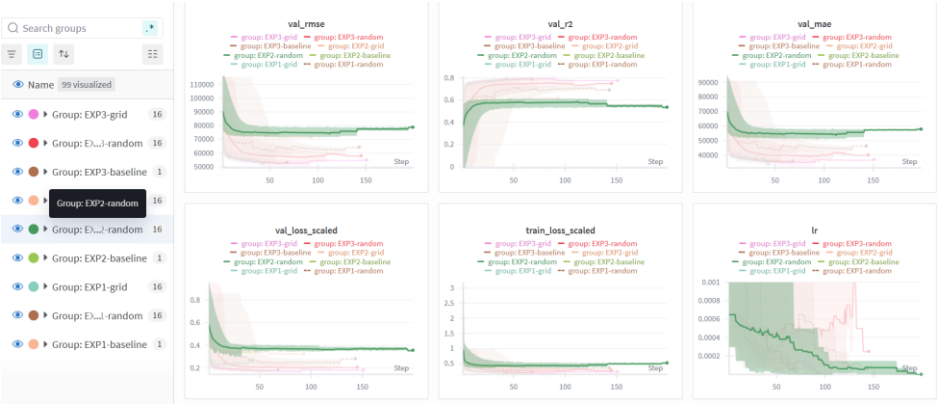
- **Train** ~71k, **Val** ~73k / ~53k MAE / R^2 ~0.60, **Test** ~75k / ~55k MAE / R^2 ~0.57
- Možeme sledovať zhoršenie vs. EXP1, malý „priestorový kontext“



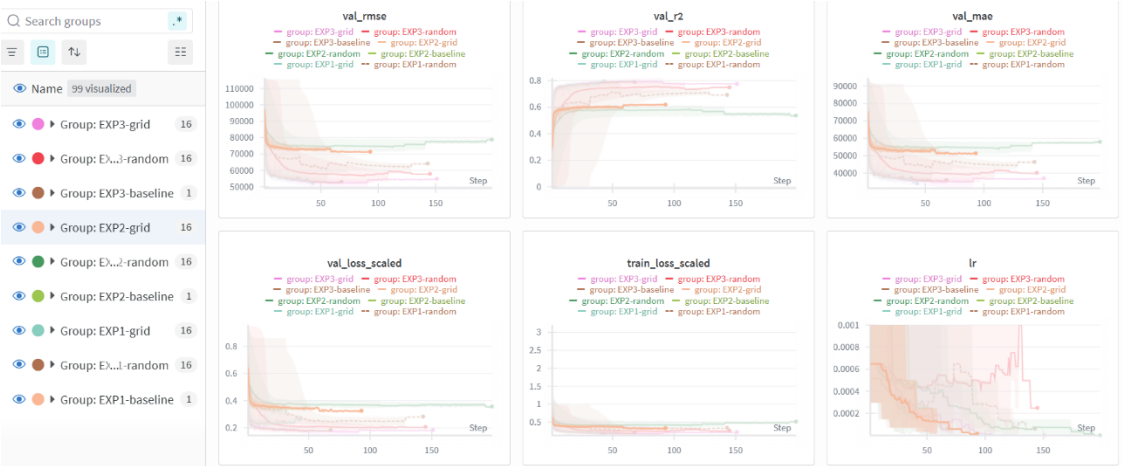


WANDB:

Random:



GRID:



EXPERIMENT 3 — Geografické bloky + štatistiky regiónov (najlepší)

Čo som spravil

- Z latitude/longitude som cez **KMeans (K≈8)** vytvoril **regióny**
- Do vstupov som pridal:
    - **one-hot** region_0..region_7,
    - regionálne **štatistiky z TRAIN**: region_mean (priemerná cena), region_size (počet vzoriek), region_rank (percentil priemeru),
    - **interakcie/polynómy**: income_lat, income_lon, lat2, lon2, lat_lon
- Celé som prehnal cez rovnaký preprocess (train-only log1p + scaler)
- Model MLP s BN/Dropout. Následne **random search** a **grid** okolo top nastavení

Najlepšia konfigurácia (grid)

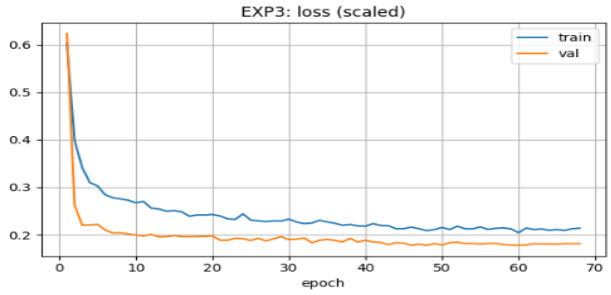
- **Layers:** [256,128,64]
- **BatchNorm:** True, **Dropout:** ~0.10
- **Optimizer:** Adam, **LR:** 1e-3, **Weight decay:** 1e-5
- **Batch:** 1024, **Patience:** ~16, **Grad clip:** 1.0
- *Feature set:* exp3_all_geo

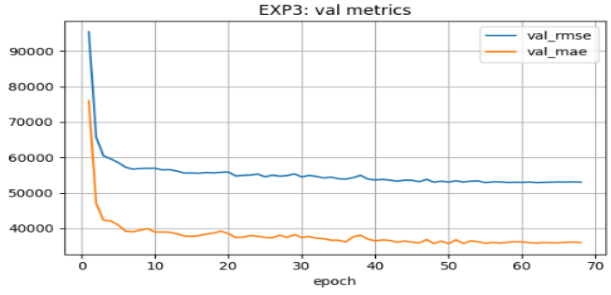
Ako to dopadlo

- Najlepšie: **Val ~49–50k, Test ~49–50k.**

Ako to pomohlo(regiony)

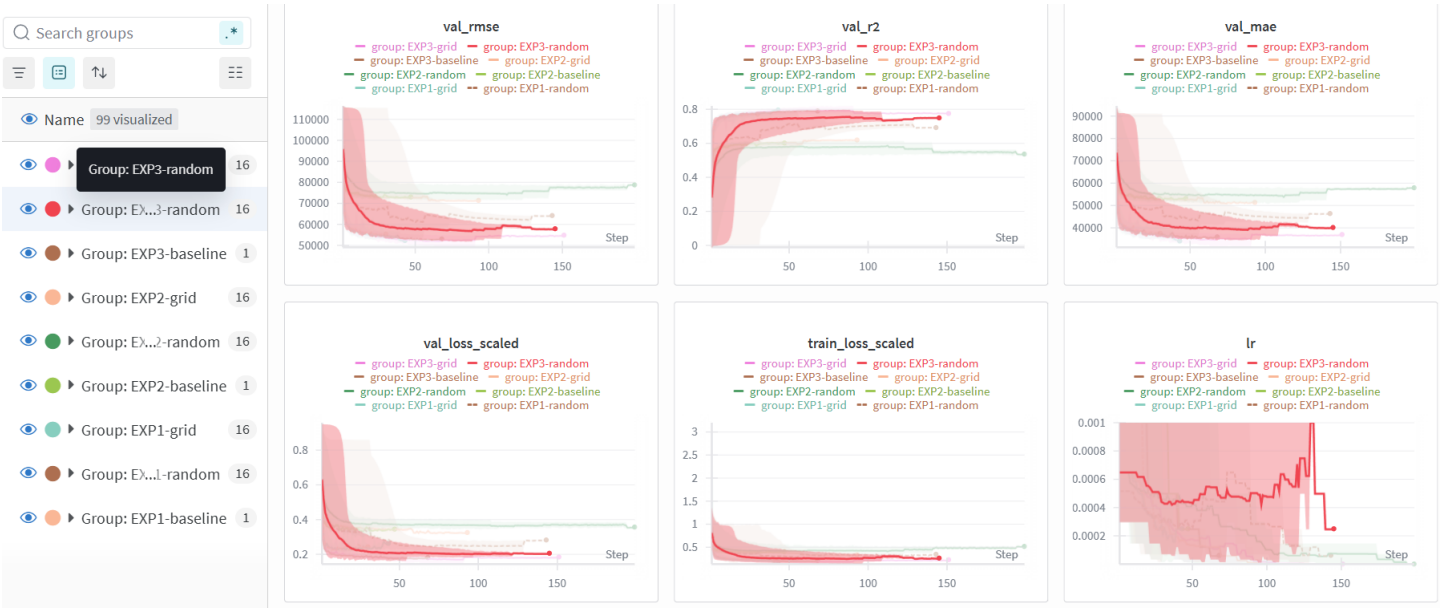
- Regióny a ich štatistiky dopĺňajú čisté súradnice o „lokálny kontext trhu“ (priemy, veľkosť klastru, ranking)
- Interakcie income × (lat, lon) zachytia gradienty cien po mape



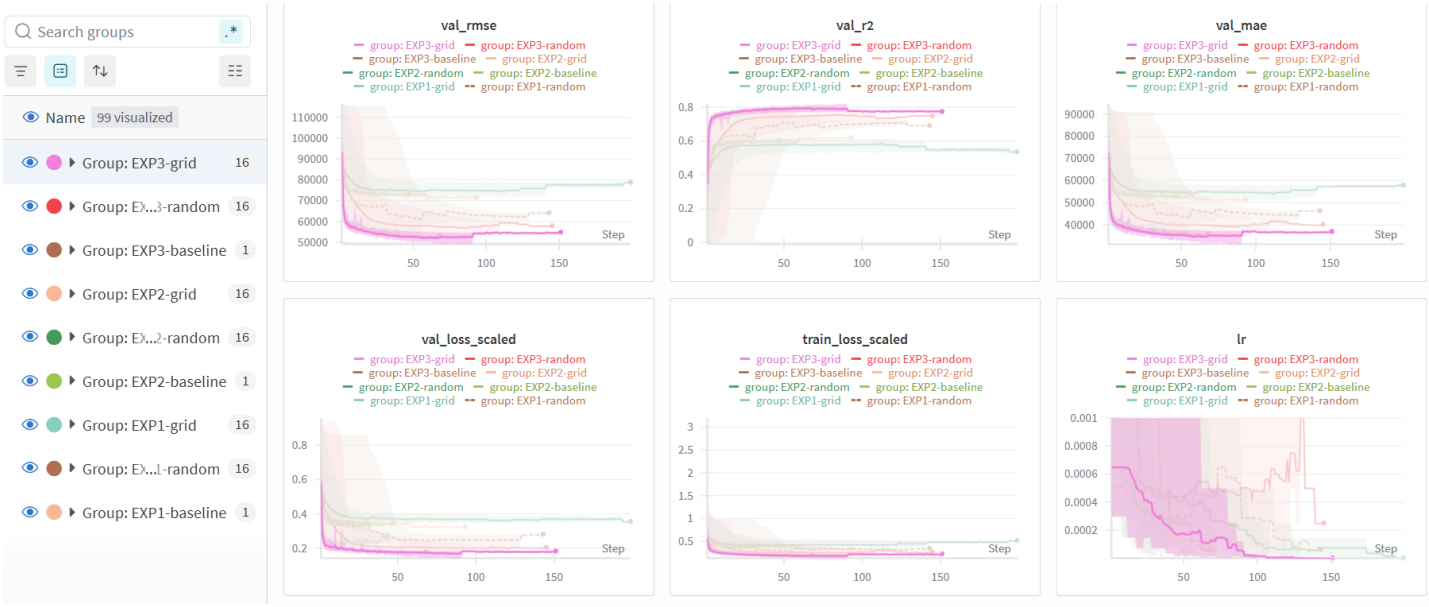


WANDB:

Random:



GRID:



4) Porovnanie experimentov

EXP1 (random)

	trial	val_rmse	val_mae	val_r2	test_rmse	test_mae	test_r2
0	8	54791.94	36329.73	0.78	55026.29	36033.54	0.77
1	13	56024.22	38416.66	0.77	56216.36	38533.17	0.76
2	4	56549.23	37687.66	0.76	56517.07	37022.33	0.76
3	5	56928.77	39401.02	0.76	57257.67	39558.74	0.75
4	6	57498.70	38930.01	0.75	57413.57	38688.31	0.75
5	2	57740.38	39894.01	0.75	57926.43	39826.64	0.75
6	12	57604.21	39773.11	0.75	58262.83	39658.39	0.74
7	3	57445.81	39754.99	0.75	58305.04	39170.13	0.74
8	9	58617.56	40584.36	0.74	58840.70	40284.08	0.74
9	10	59225.02	41836.99	0.74	59555.50	41653.15	0.73
10	1	59660.49	42130.79	0.73	59867.79	41748.04	0.73
11	14	60425.13	42679.32	0.73	61060.25	42353.48	0.72
12	16	61474.27	43886.78	0.72	61827.71	43887.54	0.71
13	7	63054.42	44005.96	0.70	63059.09	43192.93	0.70
14	15	64632.85	46823.05	0.69	65196.39	47210.70	0.68
15	11	115792.25	90481.38	-0.00	114748.84	89724.23	-0.00

EXP2 (random)

	i	val_rmse	test_rmse	val_mae	test_mae	cfg
12	13	71835.086	74327.703	51693.031	53171.566	{'seed': 9065, 'hidden_layers': [256, 128, 64,...
7	8	72054.359	74699.164	51976.797	53540.156	{'seed': 1313, 'hidden_layers': [256, 128, 64]...
10	11	72479.062	74782.102	52424.629	53674.266	{'seed': 3493, 'hidden_layers': [256, 128, 64,...
3	4	72665.438	74784.359	52597.496	53724.598	{'seed': 3007, 'hidden_layers': [256, 128, 64]...
5	6	72186.891	74867.141	52246.438	53532.004	{'seed': 6210, 'hidden_layers': [256, 128, 64]...
11	12	72363.711	74871.070	51527.621	53107.270	{'seed': 8901, 'hidden_layers': [256, 128, 64,...
14	15	72876.375	75064.609	52569.480	53867.004	{'seed': 7240, 'hidden_layers': [256, 128, 64,...
6	7	73293.875	75098.992	53534.516	54463.727	{'seed': 3567, 'hidden_layers': [256, 128, 64]...
2	3	72951.438	75746.203	53243.148	54508.277	{'seed': 1613, 'hidden_layers': [256, 128, 64]...
8	9	72767.305	75867.297	52201.207	54089.820	{'seed': 6869, 'hidden_layers': [256, 128, 64]...

EXP3 (grid)

	i	val_rmse	test_rmse	val_mae	test_mae	val_r2	test_r2
0	1	49233.66	49306.96	31749.07	31401.44	0.82	0.82
1	3	50715.01	50760.66	33081.49	33030.86	0.81	0.80
2	5	51744.96	51395.43	34778.08	34327.30	0.80	0.80
3	4	52063.84	52132.87	34464.29	34116.51	0.80	0.79
4	7	51887.34	52412.87	34846.62	34927.98	0.80	0.79
5	11	52925.28	52535.75	35500.35	35339.98	0.79	0.79
6	8	53047.28	52831.57	35405.67	35539.93	0.79	0.79
7	13	53262.09	53074.62	36198.43	36407.70	0.79	0.79
8	2	53228.40	53156.32	35860.56	35745.72	0.79	0.79
9	15	53242.44	53259.51	35997.20	36169.20	0.79	0.78

```
Najlepší grid podľa test_rmse:
 {'seed': 43, 'hidden_layers': [256, 128, 64], 'batchnorm': True, 'dropout': 0.05, 'residual': False, 'bottleneck': True, 'optimizer': 'adam', 'lr': 0.0
01, 'weight_decay': 1e-05, 'batch_size': 1024, 'epochs': 200, 'patience': 16, 'grad_clip': 1.0, 'feature_set': 'exp3_allnum_geo'}
```

5) Rozhodnutie o finálnom modeli

EXP3 (grid-tuned MLP) som vybral ako finálny, pretože konzistentne dosahuje najnižšie validačné aj testovacie chyby a najvyššie R^2
Dôležitým faktorom je **stabilita** (val ≈ test), dobré krivky konvergencie a rozumná jednoduchosť nasadenia (MLP bez ťažkých embedov)

6) Reprodukcia a technické poznámky

- **Seed:** 42 (a pevné seedy pre NumPy/PyTorch)
- **Fit len na TRAIN:** všetky transformácie (log1p, imputácia, scaler)
- **Saving:** model/artefakty len ak je to povolené (v mojom finále som ich neukladal, W&B slúži ako história)
- **Hardvér:** GPU (AMP), batch 1024
- **Tréning:** Early stopping + ReduceLROnPlateau, grad clipping 1.0, patience ~16

7) Záver

- Najväčší skok priniesli **geografické reprezentácie** (EXP3)
- Finálny EXP3-MLP (grid-tuned) má **~49–52k RMSE** na VAL/TEST pri **MAE ~34–36k** a **R^2 ~0.75–0.80**
- Krivky a metriky naznačujú **dobrú generalizáciu** a nízke riziko overfittingu