Predikce a vizualizace metorologických dat ze střední Evropy

Petr Košťál
ČVUT–FIT
kostape4@fit.cvut.cz

12. ledna 2025

1 Úvod

Cilém této práce je vytvořit model pro upřesnění predikce meteorologických dat na základě GFS[1] a datech naměřených na meteorologických stanicích. GFS je numerický model pro globální predikci počasí s přesností 13 km, který je obnovován vždy po šesti hodinách. Vypočtené hodnoty v místech stanic je možné následně extrapolovat a dostat předpověď pro celou oblast. Pro vizualizaci meteorologických dat jsem vytvořil jednoduchou webovou aplikaci v Dashi[6].

2 Vstupní data

Data jsem obdržel od Datové laboratoře (DataLab). Obsahují celkem 213 meteorologických stanic po celé České Republice a Slovensku. Problematické však je, že v nich chybí mnoho hodnot a každá stanice měří jiné veličiny, v jiných časech. Občas se také stává, že některé senzory vrací nepravdivé hodnoty z důvodu porchy. Data jsem tedy agregoval a vybral jsem nejlepší příznaky a stanice pro predikci, které nám umožní co nejobětivněji vyhodnotit přenost jednotlivých modelů. Při výběru jsem také musel zhodnotit vhdnost dat pro přislušné modely. Z tohoto důvodu jsem musel například vyřadit srážky, protože jsou většinou nulové, což je problematické pro mnoho modelů. Nakonec jsem si pro predikci vybral *teplotu vzduchu*, *přízemní teplotu* a *relativní vlhkost*. Tyto veličny měří většina stanic a neobsahují mnoho chybných hodnot.

3 Definice problému

Jednotlivé měření i předpovědi probýhají vždy šest hodin od sebe. Našim cílem bude tedy přepovědět počasí v době čtyř kroků (24 hodin) od akutálního měření.

[1] Global Forecast System

3.1 Vyhodnovací metriky

Přestože počet získaných dat nebyl příliš velký, rozdělil jsem je na trénovací a validační množinu. Jako metriku porovnání mezi jednotlivými modely jsem využil RMSE[2] na validační množině.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Pro každou veličinu i stanici jsem také spočítal MAE[3] a MSE[4] na validační množině z důvodu vhodnější interpretace chybovosti. Lze tedy porovnat nejen celkové přesnosti, ale i přesnost příslušného modelu pro konkrétní veličinu či konkrétní stanici. Modely můžeme porovnat navzájem i s referenčním GFS, z kterého vycházíme.

3.2 Regresní model

Při vytváření vlastního modelu byly vždy východiskem poslední naměřené hodnoty a GFS předpovědi. V rámci práce bylo otestováno několik modelů a optimalizačních technik. Jako první jsem vyzkoušel CatBoost[2] v kombinaci s optimalizačním nástrojem Optuna[3]. Dále jsem vyzkoušel knihovny strojového učení AutoGluon[1] a PyCaret[4], které mi efektnivně umožnili vyzkoušet mnoho různých modelů a jejich kombinace. V neposlední řadě jsem také natrénoval jednoduchou MLP[5].

3.3 Predikce časové řady

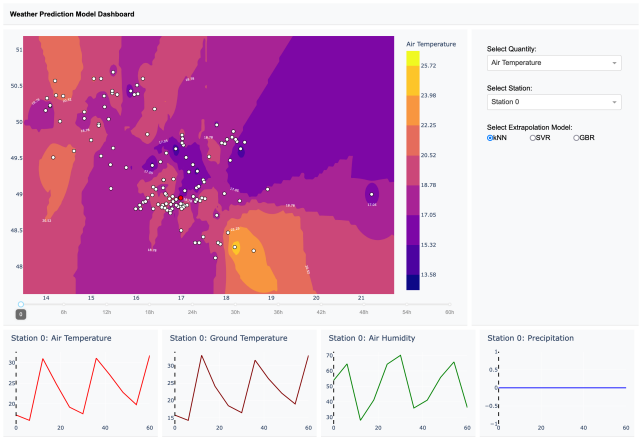
Při předpovědi časové řady jsem vždy využil poslední čtyři kroky pro předpověď následujících čtyř kroků. Model jsem trénoval tak, aby predikoval celou časovou řadu. K validaci byl však opět použit pouze výstup v čase za 24 hodin. Vyzkoušel jsem predikci pomocí rekurentní neuronové sítě LSTM[6].

[2] Root Mean Square Error
[3] Mean Absolute Error
[4] Mean Square Error
[5] Multilayer perceptron
[6] Long short-term memory

Nakonec jsem vyzkoušel model Mamba[5], který se dá rovněž aplikovat na predikci časových řad.

3.4 Ensemble modely

Nakonec jsem vyzkoušel oba postupy dohromady, tedy nejprve vytvoření modelu pro předpověď časové řady a následné natrénování modelu, který se pokusí výstup z této časové řady upřesnit. Vyzkoušel jsem kombinace LSTM a CatBoostu.



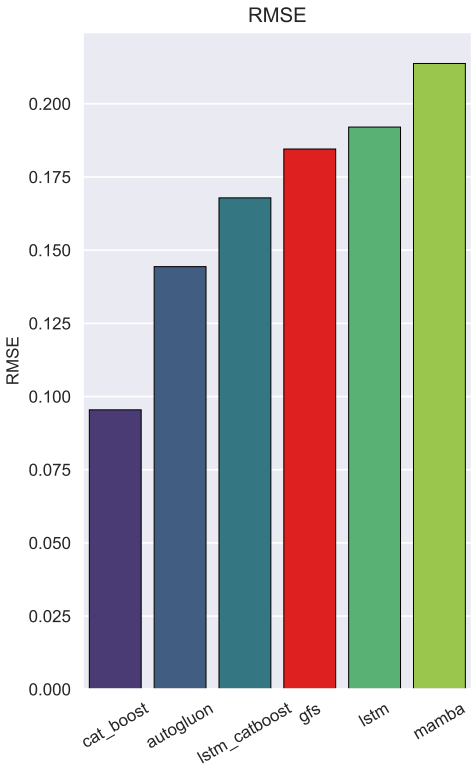
Obrázek 1: Webová aplikace vytvořená v Dash

4 Vizualizační nástroj

Vytvořil jsem jednoduchý vizualizační nástroj (obrázek č.1), který umožní data názorně zobrazit. Jedná se o konfigurovatelnou webovou aplikaci, které stačí předložit data v požadovaném formátu a vhodně poupravit konfiguraci. Nástroj umožňuje zobrazit data v jednotlivých časových úsecích i pro jednotlivé meteorologické stanice. Dále je také možné zvolit způsob extrapolace dat, v defaultním případě využijeme regresor využívající metodu nejbližších sousedů. Konfigurační soubor již obsahuje definované vhodné parametry pro demonstrační dataset.

5 Výsledky

Dle metriky RMSE na validační množině (obrázek č. 2) nám nejpřesněji vyšel model CatBoost, implementován jako regresní model. Napříč všemi veličinami (obrázek č. 3) je v průměru předpověď přesnější než referenční GFS. Druhý nejpřesnější model je AutoGluon, který rovněž konzistentně zvyšuje přesnost předpovědi. Třetí nejpřesnější model je ensemble LSTM a CatBoostu, který v průměru vychází lépe než GFS, ale průměrnou ztrátu u přízemní teploty má vyšší. Další modely spíše ztrácí oproti referenci.



Obrázek 2: Průměrná RMSE ztráta na validační množině pro normalizovaná data

6 Závěr

V rámci semestrální práce se mi podařilo vytvořit předpovědní model, který zvýší přesnost meteorologické předpovědi pro konkrétní veličiny. Přestože jsem byl hodně omezený výpočetní kapacitou, což některé modely výrazně penalizovalo, jsem s výsledkem spokojený. Zpracování témata mě bavilo a vyzkoušel jsem si mnoho nových metod strojového učení.

Reference

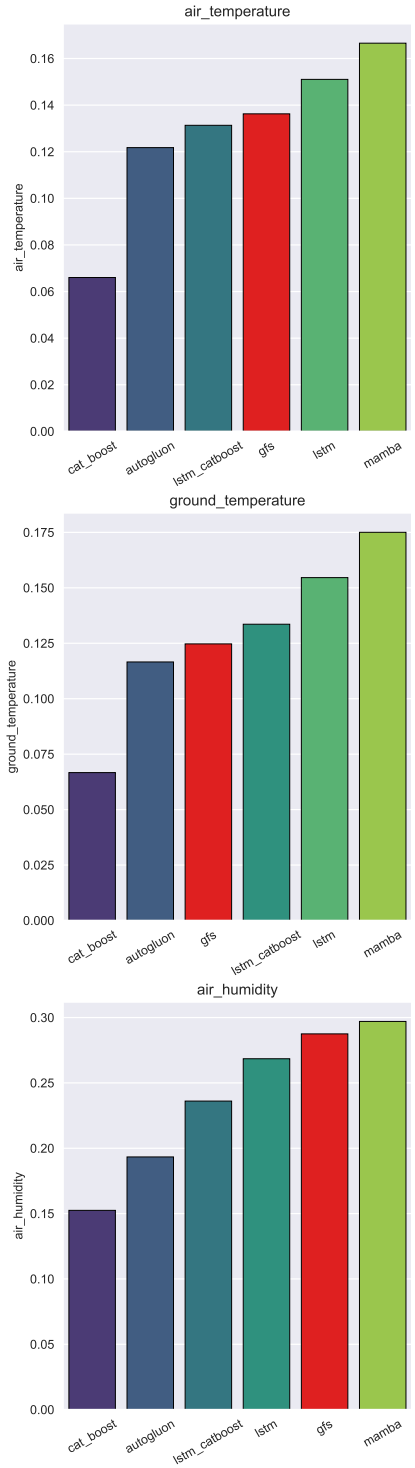
[1] Autogluon: Automl for text, image, and tabular data. online. [cit. 2025–11–1] `https://auto.gluon.ai/stable/index.html`.

[2] Catboost: a library for gradient boosting on decision trees. online. [cit. 2025–11–1] `https://catboost.ai/docs/en/`.

[3] Optuna: A hyperparameter optimization framework. online. [cit. 2025–11–1] `https://optuna.readthedocs.io/en/stable/`.

[4] Pycaret: Low-code machine learning library in python. online. [cit. 2025–11–1] `https://pycaret.gitbook.io/docs`.

[5] Tri Dao Albert Gu. Mamba: Linear-time sequence modeling with selective state spaces.



Obrázek 3: Průměrná RMSE ztráta na validační množině pro normalizovaná data pro jednotlivé veličny

online, 2023. [cit. 2025–11–1] `https://arxiv.org/abs/2312.00752`.

[6] Plotly. Dash by plotly. online. [cit. 2025–11–1] `https://dash.plotly.com/`.