Shifts Challenge Weather Prediction "CabbeanWeather" Team solution

Stepan Andreev, Master's student Surgut State University
andreev_sv@edu.surgu.ru,
Andrey Elnikov, Professor Surgut State University
andryeln@mail.ru

Abstract

This article presents an approach to solving the problem of the NeurIPS 2021 Shift's Challenge Weather Prediction track which took 2nd place in competition. The solution is based on the following main phases: data preparation, prediction of the target variable using an ensemble of catboost models, estimation of uncertainty in the forecast using the approach proposed by the organizers, and some heuristics on top of this approach.

1 Introduction

The use of machine learning methods is widespread and is used to solve many problems related to healthcare, finance, logistics, optimization of production processes, control of equipment operation, risk assessment, etc.
However, often the quality of machine learning models can show a significant deterioration in quality when working on new data and serious degradation of quality over time, which in turn can lead to extremely undesirable consequences. Such degradation is usually associated with global changes in the system that we are trying to predict and changes in the distributions of the input parameters of the model.
The task of the contest participants was to develop models that are resistant to shifts in distributions and to identify such shifts using measures of uncertainty in their forecasts. Shifts Challenge is composed of three tasks, with each corresponding to a particular data modality: tabular weather prediction, machine translation, and self-driving car (SDC) vehicle motion prediction [1].

2 Weather Prediction task

Temperature is one of the predictors of climate change [2]. This fact necessitates its constant spatiotemporal monitoring. This is done through an extensive network of temperature measurement stations distributed around the globe.
The Weather Prediction task goal was to predict the temperature at a particular latitude/longitude and time, given all available measurements and climate model predictions.

3 Data processing

In real cases, there is a common situation when you can achieve better results by processing your data. There are many different ways to handle outliers, noise, missing values, and many opportunities to enhance your data with feature selection and feature generation. In the current competition, there were several highly correlated target value features.
In addition, those features are not only highly correlated, they are actual temperatures. The "gfs temperature sea interpolated" is measured in Celsius degrees, while "cmc 0 0 0 2 interpolated" and "wrf t2" are measured in Kelvin degrees.
With respect the above there were generated two new features:
$x1 = gfs + cmc$
$x2 = (gfs + cmc + wrf - 543.8)/3$

where gfs = gfs temperature sea interpolated, cmc = cmc 0 0 0 2 interpolated, wrf = wrf t2 Feature "x2" can be used as a prediction by itself and give RMSE equals 2.40 on train data, on the other hand "x1" have much fewer missing values and can be also useful for training

4 Temperature prediction

Catboost models were used to predict temperature. There were three different model templates for training. To distinguish I named it Purple, Aqua and Silver. The depth of all models was set to 9, the maximum number of iterations is 20000, and the learning rate is 0.45. Main differences between these models presented in Table 2

Table 2: Models description

Model Name	Features Used	Folds	Add Noise	Remove Outliers	Models trained
Purple	125	25	Yes	No	32
Aqua	125	20	Yes	Yes	5
Silver	60	30	No	No	12

Add noise columns means that noise was added to the target variable while training. In Aqua models there was some filtration based on difference between target variable and "x1" feature.
Decreasing prediction errors related to number of models are shown in Figure 1.

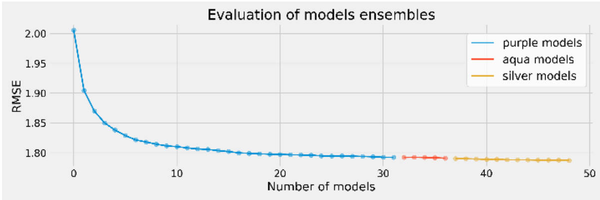

Figure 1: Models ensemble evaluation

5 Uncertainty estimation

One of the options to estimate uncertainty is to calculate predictions standard deviation across several models, lets denote such uncertainty estimations as STD.
The second opportunity is to use catboost implementation with RMSEWithUncertainty 1 objective function to predict both: target value and uncertainty estimation. There was a total of 80 models trained with this objective.
Function ensemble uncertainties regression provided by competition organizers were used to combine 80 different estimations of uncertainty, let's denote such combination result as UNCERTAINTY.
In this work, several variants have been tried to combine STD and UNCERTAINTY scores:
1. Mix UNCERTAINTY (normalized by median UNCERTAINTY) with STD (normalized by median STD) with multiplication by rate [0: 2]
2. Mix log1p from UNCERTAINTY (normalized by median log1p from UNCERTAINTY) with STD (normalized by median STD) with multiplication by rate [0: 2]
3. Mix log1p from log1p from UNCERTAINTY (normalized by median log1p from log1p from UNCERTAINTY) with STD (normalized by median STD) with multiplication by rate [0: 2]
Results with related SCORE (R-AUC MSE) are shown in Figure 2
The best result was achieved by mixing log1p from log1p from UNCERTAINTY with STD * 0.443. Denote such mixed values as UNCERTAINTY_STD_Mix.
As another way to estimate uncertainty, an heuristic based on features distribution differences between train and inference data were tested. There were selected 27 top features according to catboost model feature importance. From these features, twenty six pairs were generated, and in each pair counts objects in every grid cell, the grid is formed based on quantiles from 0 to 1 with a certain step for two parameters.

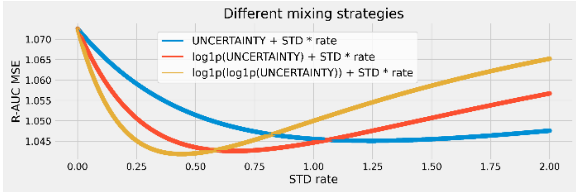

Figure 2: Mixing UNCERTAINTY with STD

Scatter plot and grid with the counted object mentioned above for pair ≪$x2$≫ and ≪cmc 0 0 0 2 interpolated≫ are displayed in Figure 3.
The number of objects in train data is determined for each object and each feature pair in test data. The Sum of that numbers can be used as an uncertainty estimator. Denote such estimator values as COUNTER.
Mixing COUNTER into Mix UNCERTAINTY STD denoted as Final UNCERTAINTY slightly improves the result. Score results on developer data with different uncertainty estimators are presented at Table 3

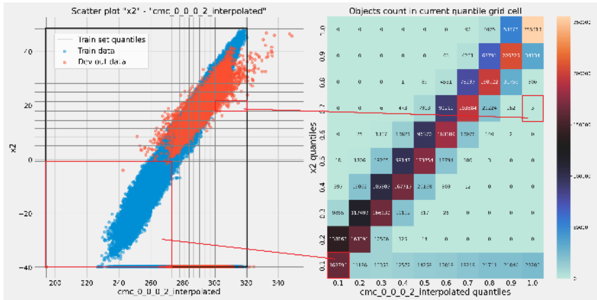

Figure 3 Grid with counted objects

Table 3: Score related to uncertainty estimator

Estimator	Score (R-AUC MSE)
COUNTER	1.361
STD	1.102
UNCERTAINTY	1.073
Mix UNCERTAINTY STD	1.042
Final UNCERTAINTY	1.039

6 Conclusion

In this paper we considered several options to estimate predictions uncertainty. We believe that some of the proposed methods can be significantly improved. Our final result was 2nd place on the final leaderboard of the competition.
Our source code can be found at https://github.com/StepanA/ShiftsChallenge

7 Acknowledgements

The work was performed within the framework of the competition "Shifts Challenge: Robustness and Uncertain-ty under Real-World Distributional Shift" of NeurIPS 2021 Competition program

References

[1] Jonathan Barichivich, Timothy Osborn, Ian Harris, Gerard Van Der Schrier, and Philip Jones. Monitoring global drought using the self-calibrating palmer drought severity index [in" state of the climate in 2019"]. Bulletin of the American Meteorological Society, 101:S51–S52, 2020.

[2] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. arXiv preprint arXiv:2107.07455, 2021.