Bike Lane Usage Forecasting Using Evolutionary Feature Construction

Hengzhe Zhang

Supervisor: Mengjie Zhang, Bing Xue, Qi Chen

Victoria University of Wellington

16/07/2023

Table of Contents

Background

Bike lane usage forecasting

- Bike lane usage forecasting predicts the utilization of bike lanes, providing valuable insights for urban planning and transportation system optimization.
- For a specific date, a lane, and associated weather conditions, the goal of bike lane usage forecasting is to develop a learning model f that is capable of predicting the usage of a given lane on a particular date.

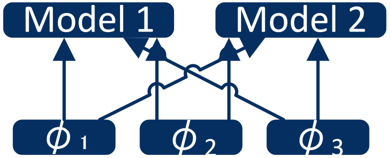
Map of Montreal

Simple idea: Build a linear regression model for each task.
Challenge:

- Linear regression may be insufficient.
- Different tasks may share some common knowledge.

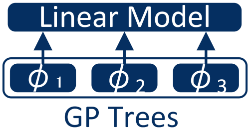
Solution:

- Construct a set of shared features for all tasks.

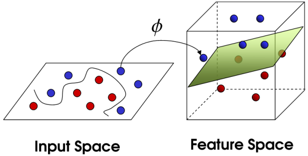


Multi-task Learning on Shared Features

- The general idea of feature construction is to construct a set of new features $\{\phi_1, \ldots, \phi_m\}$ that improve the learning performance on a given dataset $\{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$ compared to learning on the original features $\{x^1, \ldots, x^p\}$.
- Genetic programming (GP) has been widely used for automatic feature construction due to its flexible representation and gradient-free search mechanism. [1,2]



(a) Feature Construction on Linear Regression

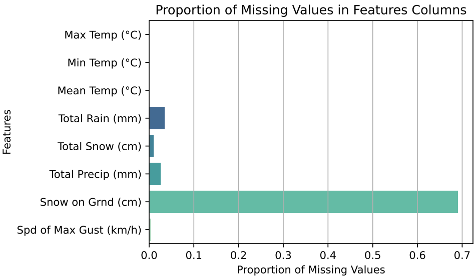


(b) New Feature Space

[1] H. Zhang, A. Zhou, et al. "An Evolutionary Forest for Regression," in IEEE TEVC, 2022.
[2] H. Zhang, A. Zhou, et al. "SR-Forest: A Genetic Programming based Heterogeneous Ensemble Learning Method," in IEEE TEVC, 2023.

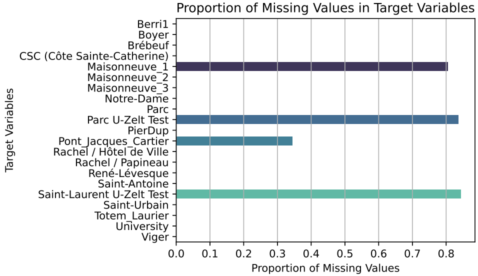
Data Preprocessing

- Missing data are imputed using the data from the previous row (yesterday).
- The column "Spd of Max Gust" contains a large number of missing values, and thus it is directly imputed with 0.



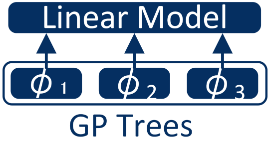
Number of Missing Values in Features

- Four target columns 'Maisonneuve_1', 'Pont_ Jacques_Cartier', 'Saint-Laurent U-Zelt Test', and 'Parc U-Zelt Test' contain missing values, and these columns are dropped, leaving 17 columns.
- However, the four dropped lanes still have some data that can be used for learning. In this paper, they are used to fit the linear model coefficients after feature construction.

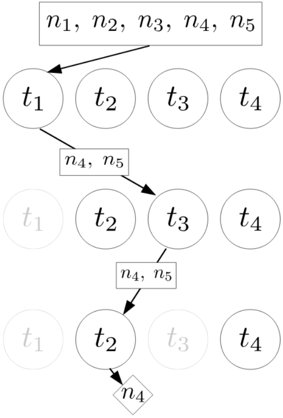


Number of Missing Values in Target Variables

Proposed Algorithm



- Population Initialization: At the initialization stage, *n* individuals are randomly initialized. Each GP individual is represented by *m* GP trees, which correspond to *m* constructed features.
- Solution Evaluation: Given the dataset with 17 lanes, 17 linear models are independently trained on the shared features for prediction. The leave-one-out cross-validation scheme is employed to enhance generalization performance.



Lexicase Selection

- Solution Selection: The cross-validation losses across all 17 tasks are denoted by $\mathcal{L}_1, \ldots, \mathcal{L}_{17}$. Traditional tournament selection is not applicable, and thus lexicase selection is employed.
- Solution Generation: New sets of features are generated based on selected individuals by applying random subtree crossover and guided subtree mutation.

Experimental Settings

Parameter Settings

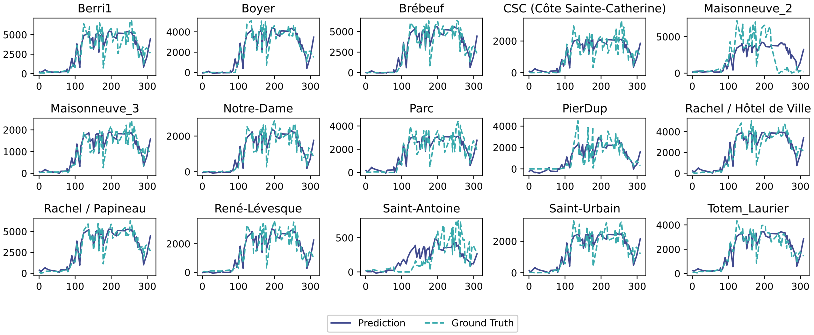
Parameter settings are common settings in GP.

Parameter	Value
Number of Generations	20
Population Size	200
Number of Trees	10
Maximum Tree Depth	3
Maximum Initial Tree Depth	0-2
Crossover and Mutation Rates	0.9 and 0.1
Functions	Add, Sub, Mul, Div

Parameter settings for all experiments.

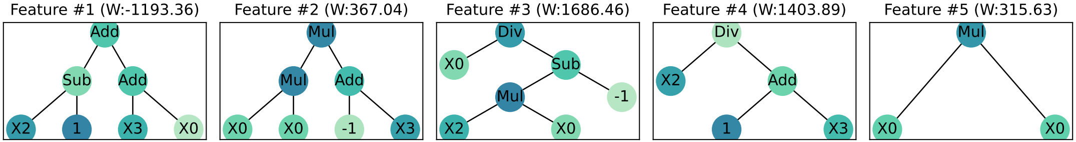
Experimental Results

- In winter, the cold weather results in fewer people opting to ride bikes.
- Conversely, the more comfortable temperatures in summer lead to a significant increase in bike ridership.
- Following a peak period in the summer, the number of bike riders gradually decreases as temperatures drop.



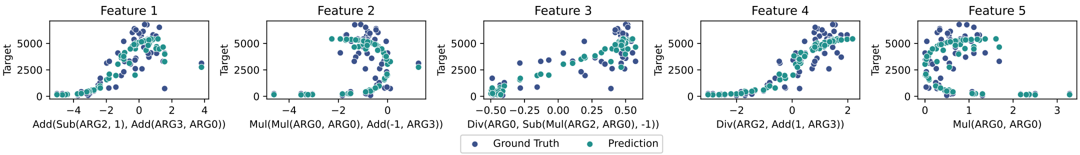
Prediction on Test Data over A Year.

- Temperature-related features x_0, x_1, x_2 and rainfall feature x_3 consistently appear in the constructed features. In contrast, snow features x_4, x_6, and gust features x_7 are absent from the constructed features.
- Snow is typically associated with winter when the temperature is low, and people are less inclined to choose to ride a bike.
- During summer, rainfall can significantly influence people's decision to ride a bike.



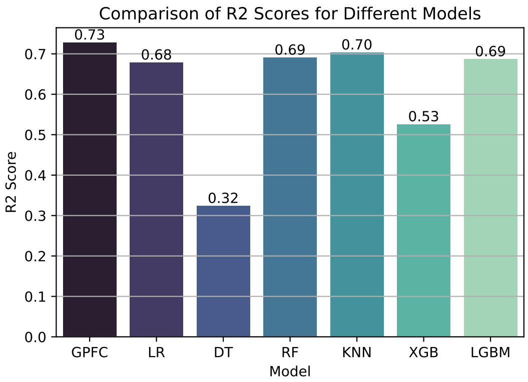
Features Constructed by GP

- $Div(ARG2, ADD(1, ARG3))$ exhibits a strong correlation between its values and the ground truth values, representing an interactive effect of average temperature and total rainfall on bike ridership.
- The combination of heavy rain and low temperature significantly reduces the number of people choosing to ride bikes.



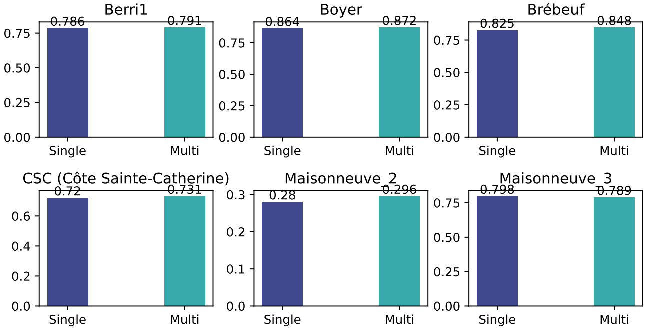
Predicted Values versus Ground Truth Values

- The GPFC achieves an impressive R^2 score of 0.73, outperforming LR, which achieves a score of 0.68.
- RF and XGB only achieve an R^2 score of 0.69.



Comparison of R^2 Scores Among Different Models

- The multi-task paradigm outperforms the single-task paradigm in 5 out of 6 cases.



Comparison on R^2 Scores of Feature Construction using Multi-Task and Single-Task Paradigms

Conclusions

- Feature construction on linear regression yields strong performance in lane usage prediction.
- Constructing a shared set of features for various tasks outperforms feature construction for each task.

Open Source Project: Evolutionary Forest (90 Stars)

Thanks for listening!

Email: Hengzhe.zhang@ecs.vuw.ac.nz

GitHub Project: https://github.com/hengzhe-zhang/EvolutionaryForest/