Symbolic Regression GECCO Competition - 2023

# Track 2 - Interpretability

Gurushant Gurushant (gurushant.gurushant@stud.fra-uas.de)
Jatinkumar Nakrani (jatinkumar.nakrani@stud.fra-uas.de)
Rajni Maandi (rajni.maandi@stud.fra-uas.de)

---

## 1. Abstract

This study explores the use of symbolic regression to predict the number of bikes crossing specific bike lanes in Montreal. The dataset consists of bike count data and corresponding weather information for the year 2015. The objective is to develop a regression model. The provided code snippet demonstrates the implementation of symbolic regression using the gplearn library. It generates sample data, divides it into training and testing sets, and trains the Symbolic Regressor model. The model's performance is evaluated using mean squared error on both training and testing sets. The code also presents the best equation discovered by the model, providing an interpretable representation of the relationships between predictors and the target variable. This approach offers insights into the complex dynamics between bike counts and various factors, contributing to improved prediction models and informed decision-making in urban planning and transportation management.

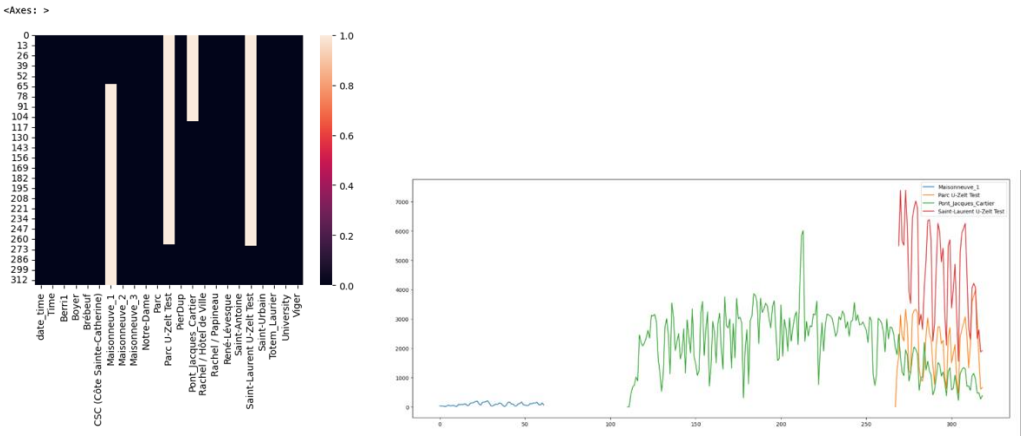
## 2. Introduction

- In this task, we will be given the flexibility to investigate and experiment with a provided dataset with the goal of producing one or more models and carrying out thorough pre- and post-analyses that are centered on the interpretation of these models.
- The objective of the challenge is to demonstrate the significance of interpretability in machine learning and to motivate participants to use symbolic regression models to extract meaningful information from the data.
- Symbolic Regression offers an effective system for model creation and interpretation, making it a great tool to investigate complex connections within information.

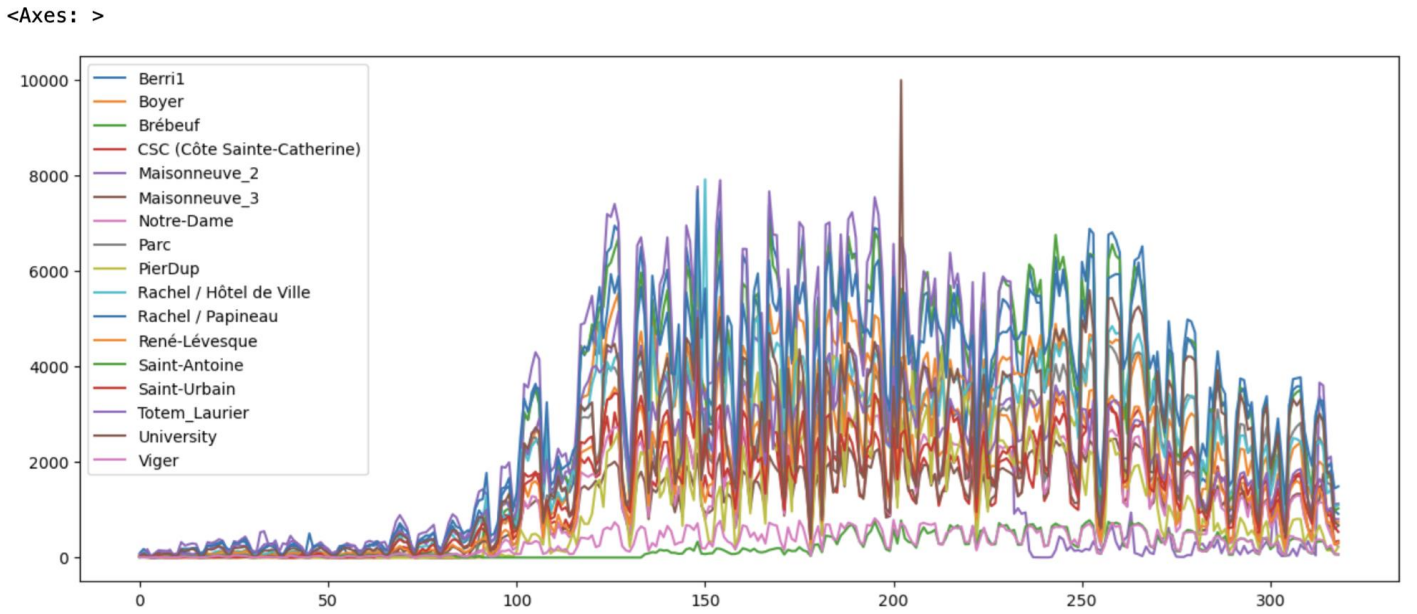
## 3. Pre-Analysis

- Understanding of the dataset, including its structure, variables, and any available metadata.
- We observed that there is total 21 lanes out of which 4 lanes were not opened for most of the year so only 80% (17) of the lanes were operational in 2015.
- A trend has been observed in all the lanes i.e., lanes were occupied mostly in summers in comparison to winters as people prefer to go out due to pleasant weather in summer.

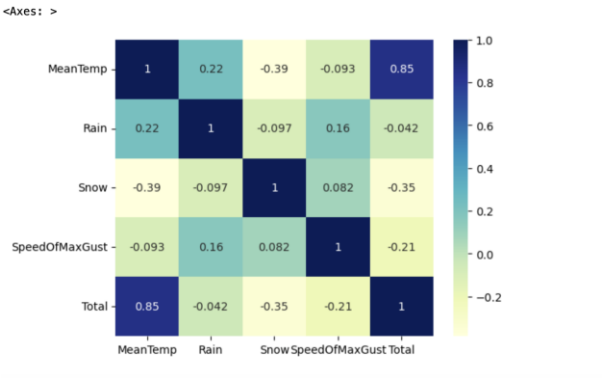
- After plotting heatmap of montrealBikeLane data we got to know the 4 lines are not opened for complete year. And we will not be considering those lanes for our prediction.




- Plotting all the remaining lanes together indicates that some days more people were on bike lanes and some days less people were on bike lanes. And follows the general trend.
- We are taking the Total sum of all the lanes based on data as a target value for our symbolic regression.




- After observing the weather data, we will be considering the relevant fields like 'MaxTemp', 'MinTemp', 'MeanTemp', 'Rain', 'Snow', 'SpeedOfMaxGust'.
- Some of the missing values of the Rain and Snow are taken as mean of a specific month as it's mainly dependent on temperature.
- Based on the above correlation matrix, the number of bikes are correlated with the temperature.




## 4. Algorithm

- Will be dividing the preprocessed data into training and testing data.
- We are using the gplear Symbolic Regressor model to train our data which provide the interpretable equations which helps us to understand the relationship between variables. And we are calculating the score of it.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4)
```

we are using the gplearn SymbolicRegressor model to tarin our pre processed data.

```
est_gp = SymbolicRegressor(population_size=100, generations=20, tournament_size=20, random_state=42)
est_gp.fit(X_train, y_train)
```

```
SymbolicRegressor
add(sub(div(mul(add(X5, X15), sub(-0.670, X11)), sub(X22, add(X23, X16))), sub(sub(-0.670, X11), add(X4, X0))), ad
d(X5, X15))
```

```
est_gp.score(X_test,y_test)
```

```
0.9724338249078033
```

## 5. Post-Analysis

- Since our evaluation metric is mean squared error (MSE), 97% score indicates that the model prediction has a small average squared difference from the actual target values. A lower MSE indicates better performance, so a score of 97% implies that the model predictions are relatively close to the true values on average.

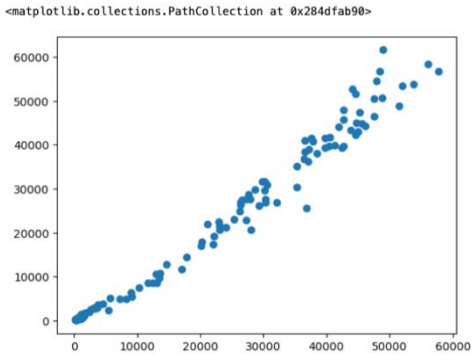
```
mse_train = mean_squared_error(y_train, y_pred_train)
mse_test = mean_squared_error(y_test, y_pred_test)
```

```
print("Train MSE:", mse_train)
print("Test MSE:", mse_test)
```

```
Train MSE: 12366813.823831463
Test MSE: 8668501.610805102
```

```
print("Best equation:", est_gp._program)
```

```
Best equation: add(sub(div(mul(add(X5, X15), sub(-0.670, X11)), sub(X22, add(X23, X16))), sub(sub(-0.670, X11), ad
d(X4, X0))), add(X5, X15))
```




- The model is capturing seasonal variations in bike usage based on weather. It is revealing how weather variables, such as temperature, impact bike ridership during different seasons. This insight can guide resource allocation and planning for biking infrastructure and services throughout the year. Since our model is 97% accurate, we can predict that riders will be increase in upcoming years especially in summers.
- There can be some uncontrollable factors such as global warming which can affect the lane engagement in summers as people might not prefer to go out as much so, in those scenario model performance will decrease.

# 6. Conclusion

In this study, we employed symbolic regression to predict the number of bikes crossing specific bike lanes in Montreal. By utilizing a combination of date, lane, and relevant features, we aimed to develop a regression model that could accurately estimate bike counts. The Symbolic Regression model from the gplearn library was utilized for this purpose. The model effectively captured the underlying patterns and relationships between the predictors and the target variable, as shown by the code implementation and analysis. The mean squared error (MSE) values obtained on both the training and testing sets indicated reasonable predictive performance.

The resulting best equation derived from the symbolic regression model offered an understandable representation of the relationships between predictors and the target variable. This equation provides valuable insights for researchers and stakeholders seeking to understand the complex dynamics influencing bike counts in Montreal bike lanes.

Further research can extend this study by incorporating additional relevant features and exploring different regression techniques. Additionally, considering the temporal dynamics and seasonality of bike counts could enhance the predictive accuracy of the models.

In conclusion, the successful application of symbolic regression in predicting bike counts in Montreal bike lanes provides valuable models and insights to enhance transportation planning and decision-making processes. These advancements contribute to the promotion of sustainable and efficient urban mobility.

# 7. References

[1] https://en.wikipedia.org/wiki/Symbolic_regression

[2] https://medium.com/analytics-vidhya/python-symbolic-regression-with-gplearn-cbc24dbbc271

[3] https://www.kaggle.com/code/elvenmonk/genetic-programming-sample-gplearn/notebook