Hourly Bike Share Check-Outs and Check-Ins Prediction in Toronto for Top 15 Stations with Most Data

1. Abstract

This project focuses on predicting hourly ridership checkouts and check-ins for Toronto's bike share system using data from 2019 to 2024. The goal is to forecast ridership for March 2024 as the ridership data for March 2024 is not available at the time of this project. The ridership data was obtained from Toronto's open data portal, while the weather data is obtained from the government of Canada's site. The analysis involves data preprocessing to scope the dataset to the top 15 stations with the most data, with which the application of machine learning techniques are used to find the best performing predictive model. The model is trained on data from 2019 to 2022, validated on the 2023 dataset to find the best model and its hyperparameters. Then, the final model is trained with the data inclusive of 2023 and used to predict January-February 2024 data to calculate for accuracy, and then finally prediction is generated for the period of March 2024 for further analysis.

2. Introduction

Bike-sharing systems have become an integral part of urban transportation networks, offering a sustainable alternative to traditional modes of transport. Predicting ridership can help optimise the distribution of bikes across stations, reduce operational costs, and enhance user experience. This project leverages machine learning techniques to find a model that best predict hourly bike share ridership in Toronto using historical ridership and weather data.

3. Data

3.1. Data Acquisition

The data used in this project includes:

- **Ridership Data (2019-2024):** This data details the anonymised trip data in a per trip basis at various stations across Toronto, was obtained from the Toronto Open Data Portal. The data was downloaded and unzipped for each year from 2019 to 2024. After unzipping, any subdirectories were consolidated to ensure all files were correctly organized by year. The data was then aggregated by hour for each station, creating a dataset to be used for analysis and prediction.
- **Weather Data (2019-2024):** This data includes variables such as temperature, humidity, wind speed, and precipitation, was obtained from the Government of Canada. This data was similarly downloaded for the corresponding years and merged with the ridership data based on timestamps,

ensuring that each hourly check-out and check-in per station had matching weather conditions.

3.2. Data Aggregation and Cleaning

Aggregation and cleaning of the data is performed via the following:

- **Aggregating Ridership Data:** The ridership data was aggregated by hour for each station to create a data that can be used to depict the check-ins and check-outs over time. This aggregation was necessary to align the ridership data with the corresponding hourly weather data.
- **Merging Datasets:** The aggregated ridership data was merged with the weather data using a left join based on timestamps, ensuring that each ridership record had corresponding weather conditions.
- **Handling Missing Values:** Missing values in the dataset were addressed using forward fill and dropping of datapoints with NA values.
  - **Forward Filling:** Temporal data gaps were filled using forward filling, which propagates the last known value to subsequent missing entries. This technique was specifically used for the "Weather" column of the data.
- **Binarise and Aggregate Weather Categories:** The "Weather" column in the combined dataset was binarised into its seperate categories (e.g., "Rain, Snow" becomes "Rain" and "Snow" in its own column) instead of its previously combined form where multiple categories of weather can be in the same value. These binarised categories were then aggregated into more general groups (e.g., Rain-related, Snow-related, Clear/Cloudy) to reduce data dimensionality.

3.3. EDA and Data Preprocessing

Exploratory Data Analysis and further data preprocessing steps are completed, which produces the final dataset used for modelling.

3.3.1. Feature and Data Analysis

- **Initial Features:** The dataset initially consisted of 19 feature columns, including various weather variables and ridership metrics.
- **Station Data Scope:** A preliminary analysis of the number of datapoints per "Station Id" revealed that some stations did not have enough data to provide the model with sufficient information for training. Therefore, the scope of the model was limited to the top 15 "Station Id" that had the most datapoints, each with more than 24000 datapoints.
- **Data Distribution Comparison:** To validate the representativeness of the scoped dataset, the "Check-Outs" and "Check-Ins" were plotted against "Month," "Day," and "Hour" for both the full dataset and the

scoped dataset. The resulting graphs showed that the shapes of the curves were very similar, indicating that the scoped dataset could effectively represents the distribution of the full dataset.

3.3.2. Correlation Analysis

- **Correlation Matrix:** A correlation matrix was computed to identify relationships between features to help in the understanding of which features are less crucial or redundant.
- **Feature Selection:** Based on correlation analysis, highly correlated features were identified, and redundant features were considered for removal to reduce multicollinearity.

3.3.3. Dropping Correlated Features Based on the correlation analysis, the following decisions were made: - **Drop Temp (°C) or Dew Point Temp (°C):** Dew Point Temp (°C) was dropped because it had a high correlation with Temp (°C), and Temp (°C) showed a stronger correlation with the target variables. - **Low Correlation Features:** Features with low correlation to the target variables, such as Stn Press (kPa) and Wind Dir (10s deg), were dropped.

4. Modelling Methodology

4.1. Overview

The modeling process involved several key steps aimed at identifying the best predictive models for bike share ridership (both check-outs and check-ins) in Toronto. The steps were as follows:

1. **Training and Validation Datasets:**
   - The models were trained on data from 2019 to 2022 which were then validated on the 2023 dataset, which served as an unseen dataset to evaluate the initial performance of each model.
2. **Model Selection Based on R2 Score:**
   - Several models were initially tested, including Linear Regression, Ridge Regression, Support Vector Regression (SVR), Random Forest Regression, Gradient Boosting Regressor, XGBoost Regressor, and K-Nearest Neighbors (KNN).
   - For each target variable (check-outs and check-ins), the top two models were selected based on their R2 score on the 2023 validation set. The R2 score was chosen as the primary metric to assess how well the models captured the variance in the data.
3. **Hyperparameter Tuning:**
   - The top two models for each target variable underwent further refinement through hyperparameter tuning. This process was done in two stages:
     - **Randomized Search:** Initially, a randomized search was conducted to broadly explore the hyperparameter space and identify

potential optimal ranges for each model's parameters.
  - **Grid Search:** Following the randomized search, a more focused grid search was performed within the identified ranges to fine-tune the hyperparameters. This methodical approach ensured that the final models were optimised.
4. **Final Model Selection:**
   - The model with the best performance after hyperparameter tuning was selected as the final predictive model for each target variable. These models were then used to predict ridership.

5. Results

5.1. Modelling Methodology Recap

- **Evaluation Metrics:** The models were evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) on the validation set (2023 data).
- **Model Comparison:** Performance across models was compared to determine the best approach for deciding on the final two models for hyperparameter tuning.

5.2. Interim Modelling Results

The following table summarises the performance metrics for each regression model tested. The metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) for both Check-Outs and Check-Ins:

Model	Metric	Check-Outs	Check-Ins
Linear Regression	MAE	2.3504	2.5325
	MSE	12.2170	15.2007
	R^2	0.2343	0.2141
Ridge Regression	MAE	2.3504	2.5325
	MSE	12.2170	15.2007
	R^2	0.2343	0.2141
Random Forest	MAE	1.9731	2.0928
	MSE	8.4628	10.2337
	R^2	0.4696	0.4709
Gradient Boosting	MAE	1.9153	2.0456
	MSE	8.0778	10.1540
	R^2	0.4937	0.4750
Support Vector	MAE	2.7221	3.6603
	MSE	16.8811	27.1586
	R^2	-0.0580	-0.4042
K-Nearest Neighbors	MAE	2.1778	2.3345
	MSE	11.1797	13.9587

Model	Metric	Check-Outs	Check-Ins
XGBoost Regressor	R^2	0.2993	0.2783
	MAE	1.9505	2.0994
	MSE	8.2106	10.2667
	R^2	0.4854	0.4692

5.3. Final Modelling Results:

The top two models based on R2 scores—**Gradient Boosting Regressor** and **XGBoost Regressor**—will be selected for hyperparameter tuning (Random Forest Regressor had a better R2 score than XGBoost Regressor for Check-Ins, however, XGBoost Regressor was still chosen as it was only slightly poorer in its result for Check-Ins but had a much better R2 score than Check-Outs than Random Forest). The tuning process will include both randomised search and grid search methods to optimise the models.

After hyperparameter tuning, the top models for predicting Check-Outs and Check-Ins were selected based on their R2 score.

5.3.1. Check-Outs

- **XGBRegressor:**
  - **Hyperparameters:** `n_estimators=1000`, `max_depth=10`, `learning_rate=0.01`, `reg_alpha=1.0`, `reg_lambda=1.0`
  - **Performance:**
    - MAE: 1.9131
    - MSE: 8.0068
    - R^2: 0.4982
- **Gradient Boosting Regressor:**
  - **Hyperparameters:** `subsample=1.0`, `n_estimators=100`, `max_depth=10`, `learning_rate=0.1`
  - **Performance:**
    - MAE: 1.9229
    - MSE: 8.0526
    - R^2: 0.4953

5.3.2. Check-Ins

- **XGBRegressor:**
  - **Hyperparameters:** `subsample=0.6`, `n_estimators=1000`, `max_depth=6`, `learning_rate=0.1`, `colsample_bytree=0.6`
  - **Performance:**
    - MAE: 2.0493
    - MSE: 9.9365
    - R^2: 0.4863
- **Gradient Boosting Regressor:**

- **Hyperparameters:** `subsample=1.0`, `n_estimators=100`, `max_depth=10`, `learning_rate=0.1`
- **Performance:**
  - MAE: 2.0425
  - MSE: 9.8948
  - R^2: 0.4884

5.4. Selected Models for Final Predictions

Based on the hyperparameter tuning and R2 score:

- For **Check-Outs**, the **XGBRegressor** was selected as the final model due to its slightly better performance in terms of R2 (and the other metrics).
- For **Check-Ins**, the **Gradient Boosting Regressor** was chosen for its better R2 score (and other metrics as well).

These models will be used to generate the final predictions for March 2024, providing hourly forecasts for both checkouts and checkins.

6. Model Predictions

The best-performing model, after hyperparameter tuning, will be used to predict ridership for Jan-Feb 2024 (truth values were available as the ridership data for this period is published) and March 2024 (truth values were not available), providing hourly forecasts for checkouts and checkins.

6.1. Model Evaluation on Unseen Data

After selecting the final models based on their performance during the hyperparameter tuning phase, the models were tested on an unseen dataset consisting of ridership data from January and February 2024. The goal was to evaluate the models' generalisation capabilities and their effectiveness in making accurate predictions on new data.

6.1.1. Performance on 2024 Ridership Data (January and February)

- **Check-Outs:**
  - MAE: 1.2096
  - MSE: 2.8596
  - R2: 0.2894
- **Check-Ins:**
  - MAE: 1.2145
  - MSE: 2.9988
  - R2: 0.3452

6.1.1.1. Analysis of Model Performance Metrics (on Jan-Feb 2024 data)

- The **Mean Absolute Error (MAE)** and **Mean Squared Error (MSE)** indicate that the models performed reasonably on the unseen dataset, with both metrics showing acceptable levels of error. The MAE and MSE were lower than the values during training - however it could be attributed that the lower error values in this case were due to the lower true checkouts and checkins in general for this period.
- The **R2 scores** suggest that while the models were able to capture some of the variability in the ridership data, there is room for improvement. Specifically, the R2 scores for both checkouts and checkins indicate that approximately 29% and 35% of the variance in the data, respectively, can be explained by the models. This was much lower than the R2 score produced during training.

6.1.1.2. Analysis of Model Prediction (for Mar 2024 data) The checkouts and checkins models will be used to predict ridership for March 2024. The max checkouts and checkins are then compared with the capacity of each station.

Station Id	Capacity	Check-Outs	Check-Ins
7006	31	11.17	10.46
7033	43	8.82	11.42
7121	27	5.54	5.66
7100	27	8.33	10.98
7076	57	19.17	15.36
7022	39	8.09	10.03
7030	35	8.20	6.71
7007	19	7.42	6.71
7044	35	3.69	5.09
7089	27	6.47	7.45
7389	25	6.52	5.70
7078	15	7.63	9.49
7038	31	12.97	12.53
7253	12	6.48	6.19
7102	31	7.80	8.11

It appears that none of the maximum checkouts that were predicted for all stations were beyond the capacity of the station. Instead, it could be interpreted that the capacity at some stations were beyond the demand for bikes at those stations. However, it should be noted that this represents a snapshot of checkouts and checkins against the capacity at a specific point in time (i.e. the hour at which a maximum checkout and checkin had occurred). As such, the influence by the checkouts and checkins of due to the previous or future hours were not considered (e.g. using the above prediction for Station Id 7009, even though the checkouts of 11.17 is significantly smaller than the capacity of 31, the checkouts and checkins of the hour before this datapoint was not considered, which could

have reduced the capacity of the station to a value below 11.17 which would have had resulted in a lack of supply for the station).

7. Improvements

- **Using Deep Learning Techniques:** Exploration of deep learning models for improved accuracy could be considered. Related Study. Deep learning models could capture temporal dependencies and spatial patterns more effectively.
- **Additional Features:** Incorporating features like seasonality, land use, and spatial relationships between stations could enhance model performance. For example, including information about the land use of the areas nearby which could result in higher traffic flow in the area to create more demand for ridership.
- **Spatial Relationships:** Including spatial relationship between stations could aid to improve the model as the checking out of bikes at one station usually leads to a checking in of bikes at another station.

8. Conclusion

This project predicted bike share ridership for March 2024 using historical data from 2019 to 2023. Ridership data and weather data were downloaded, aggregated and combined to form a comprehensive dataset that can be used for modelling of ridership in the top 15 stations in Toronto with most datapoints. The model predicted the demand of bike share to be within the capacity of the stations, with certain stations potentially having much higher capacity than demand. However, it should be noted that there are also limitations to the model in terms of its performance metrics that might be improved with additional data and/or more complex models (i.e. neural networks).

9. References

- Toronto Open Data
- Government of Canada
- ChatGPT