Predicting Bikeshare Demand

Ben Bruce, Chutian Ma, Keith Mills, Beni Pazár, Shaoyang Zhou

Motivation

- City bikeshare programs allow individuals to check out a bike for a short period of time.
- Modeling bikeshare demand is essential for:
    - Providing reliable access to bikes
    - Scheduling maintenance and repairs
    - Planning future bike infrastructure

Objectives

- Model daily usage of Mobi, the bikeshare in Vancouver, British Columbia.
- Quantify the impact of weather and climate on the number of bike trips taken each day.
- Forecast bikeshare usage for a given date and given weather conditions.

Data Collection & Cleaning

Collection

- Bike trip data from Mobi (published monthly)
- Weather data from Environment and Climate Change Canada (published daily)
- Daylight hours from timeanddate.com

Features

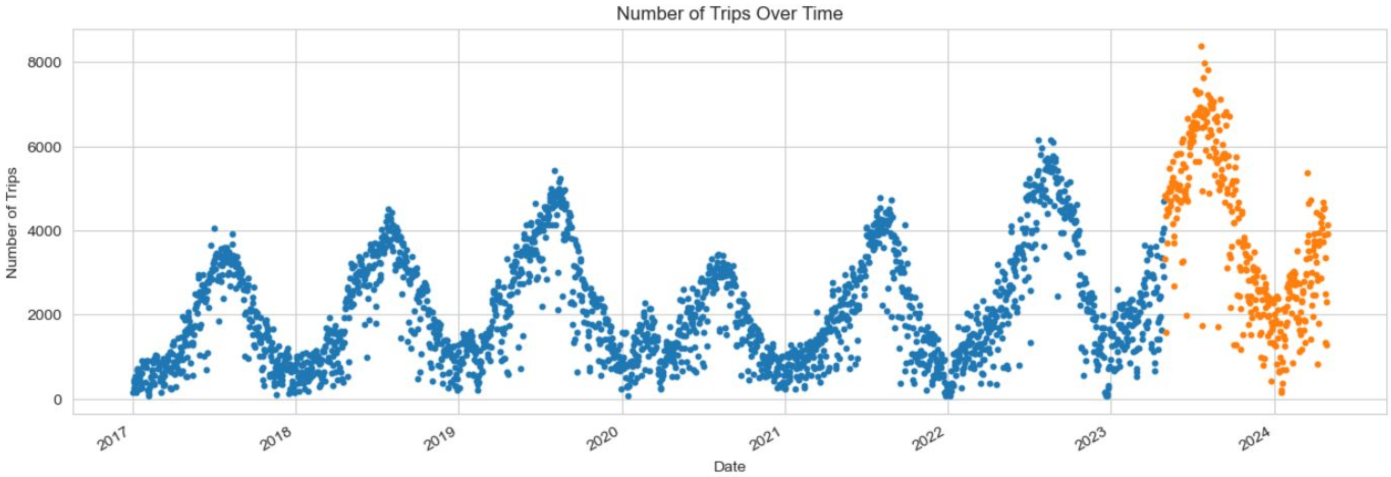
18 weather features, including `day_length`, `max_temp`, `temp_diff`, `total_precip`, `max_gust`

Cleaning

- Missing temperatures replaced using linear interpolation
- Missing precipitation replaced using averaging
- Daily and monthly data combined into a single file spanning January 2017 to April 2024

Initial Data Visualization

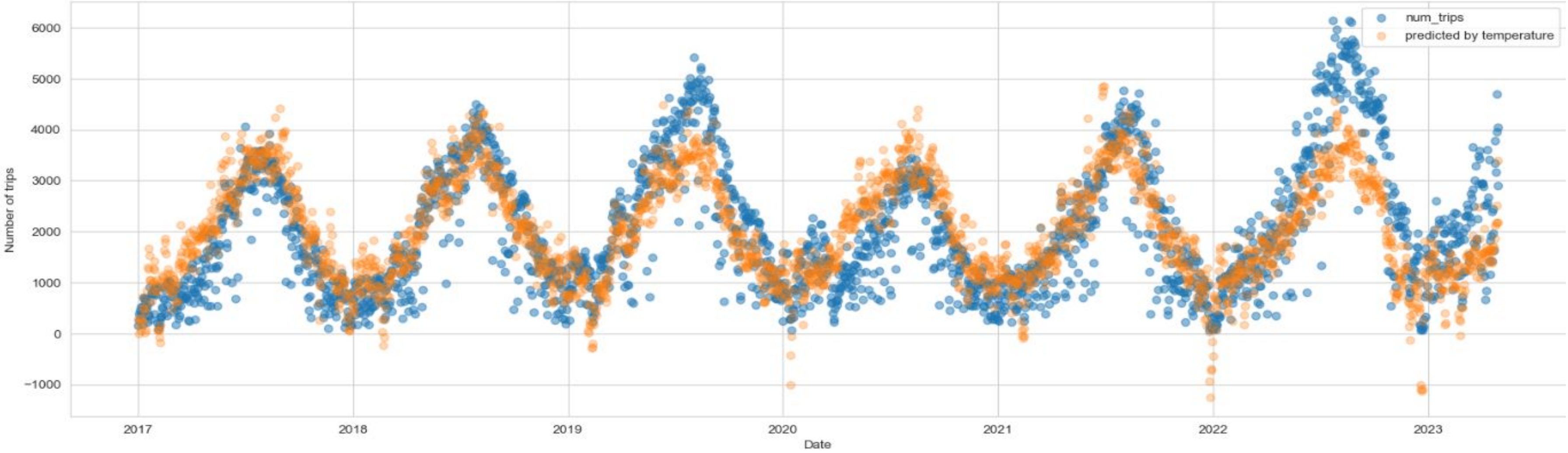
- Train-test split setting aside the last year of our data for testing (May 2023 to May 2024)



- Training data exhibits seasonality and a generally increasing trend, with a “reset” in 2020 due to COVID-19 (and faster recovery and growth afterwards)

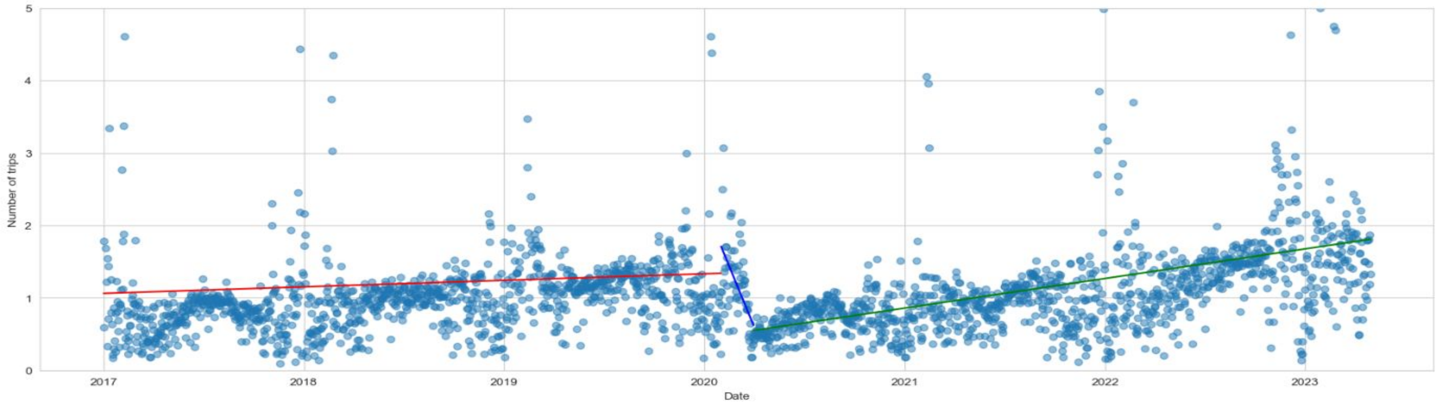
Analyzing Seasonality and Trend

- The trend and seasonality interact with each other
- We found that a multiplicative time-series model worked best (out of four total approaches)
- The seasonality component was explained by temperature:



Analyzing Seasonality and Trend (cont'd)

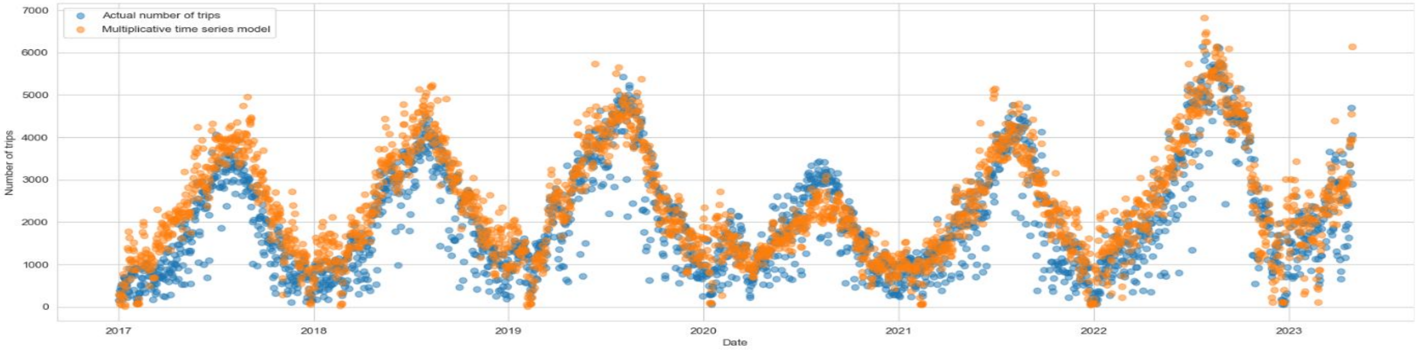
- After removing seasonality, we found a piecewise-linear trend with three parts:



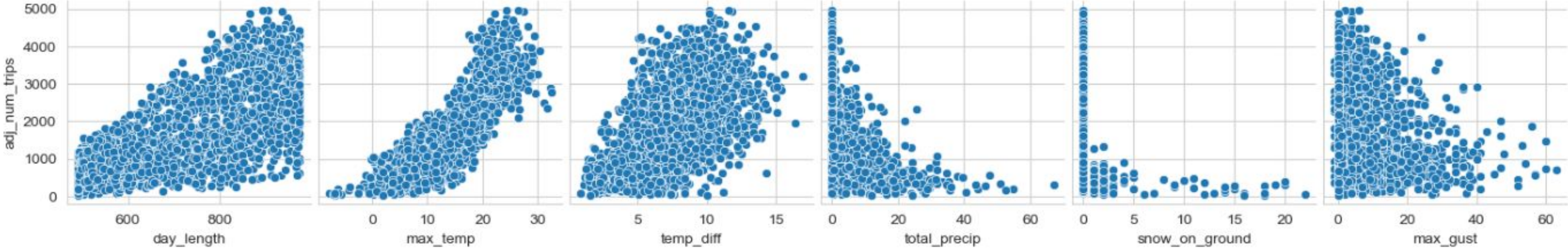
- This trend is roughly due to base demand for the bikeshare, rather than weather data
- Correcting for this trend helps keep COVID weather data useful
- We removed this trend, making an "adjusted" number of trips, and aimed to predict this number from weather data

More Data Visualization

- Our final time series model with seasonality explained by daily maximum temperature and piecewise linear demand trend



- After feature selection from various plots, correlation matrices, and lasso regression for feature analysis, the following six weather features emerged as most related to the number of bike trips taken in a day



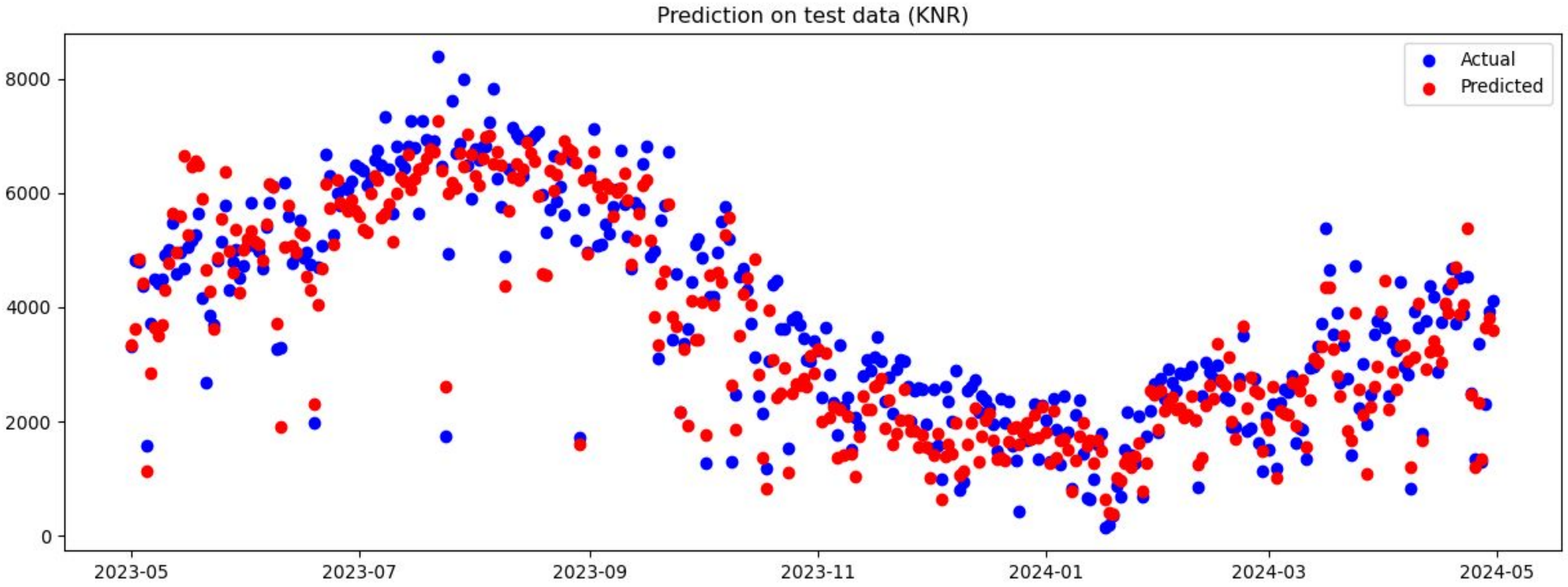
Modeling approaches

- **Splitting:** Training data: 2017 - 2023. Test data: 2023 – 2024.
- **Target:** `y = adj_num_trips` (number of trips after detrending).
- **Model categories:** Linear regressions, KNRs, Tree-based models (Decision-Tree, XGBoost), Time series
- **Feature selection & hyperparameter tuning:**
  - Perform 5-fold cross validation on the training set, computing the average root mean sqaure error (RMSE).
  - Repeat for all combinations of features/hyperparameters, and select the one with the minimal CV RMSE.
- **Compute training and testing RMSEs based on** `num_trips` (before detrending).

Model comparisons

Model	Features/Hyperparameter	RMSE		
		Training	CV	Testing
Linear Regression	`day_length`, `max_temp`, `temp_diff`, `total_precip`, `max_gust`, interaction terms	495	457	731
KNR on dates	K = 60	605	540	1010
KNR on six weather features	K = 16, Manhattan metric	420	410	677
Decision tree regressor	`max_temp` , `total_precip`	415	476	682
XGBoost	`max_temp`, `total_precip`, `day_length`	421	440	684

Performance on the test data



Model Interpretations & Implications

Why does KNR on six weather features work relatively well?

- Input values are close to those in the training set.

- The Manhattan metric is more preferable for high-dimensional data

Feature Importance (via decision tree regressors)

Top 3:

- Max temperature
- Total precipitation
- Day length

Least impactful:

- Wind speed
- Temperature difference
- Snow on ground

Future directions

- More ways of detrending the data
    - We predicted de-trended trips much better than the actual number of trips, suggesting that our trend removal can be improved
- Tailor the model for bike stations with the most demand
- Train on hourly data to get more accurate models
- Expand our model to include more cities with bikeshare programs

Thank you!

The team members express their gratitude to Dr. Roman Holowinsky, Dr. Steven Gubkin, and Dr. Alec Clott for organizing the bootcamp and giving us an amazing learning opportunity. We also thank our mentor, Dr. Dyas Utomo, for overseeing our project and providing us with various insights and comments.