Erdős Summer 24 Executive Summary

Predicting Mobi Bikeshare Usage from Weather

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

Project Description:

This project presents a predictive algorithm designed to forecast the daily number of rides taken by Mobi users. The primary objective is to enhance Mobi's ability to manage its fleet of bikes efficiently. Accurate trip predictions enable Mobi to ensure a sufficient number of bikes are available, optimizing operational efficiency and maximizing revenue. By anticipating user demand, Mobi can strategically allocate resources, ensuring bikes are available during peak times and scheduling maintenance during periods of low demand. This proactive approach not only maximizes revenue but also enhances user satisfaction by minimizing service disruptions.

Data Utilized:

The data was sourced from Mobi's online database, encompassing the period from 2017 to 2024. This comprehensive dataset was enriched with local weather data obtained from the Environment and Climate Change Canada database. Key features incorporated into our analysis include average and maximum daily temperature, total daily precipitation, hours of daylight, amount of snow on the ground, and the difference between minimum and maximum daily temperatures. The user usage data exhibited cyclic patterns with a roughly linear growth trend and varying amplitude, all of which were considered in our modeling approach.

Methodology:

The data underwent rigorous cleaning to eliminate duplicates and handle missing values. Primary preprocessing efforts focused on normalizing the data to account for growth trends and the significant disruptions caused by the COVID-19 pandemic. Various methods were considered, including linear and sinusoidal subtraction, but a multiplicative rescaling factor was ultimately employed for this purpose.

Our modeling approach included the implementation of several regression techniques: linear regression, k-nearest neighbors (KNN) regression, decision tree regression, and gradient-boosted tree regression. To determine the most effective model, we compared the root-mean-square error (RMSE) of the predicted values against the actual data.

The KNN regression model, with K=16 and using the Manhattan distance metric, demonstrated the best performance. This model incorporated features such as day length, maximum temperature, temperature difference, total precipitation, amount of snow on the ground, and maximum wind speed. The KNN regression model achieved an RMSE of 420 on the training data and 677 on the testing data, indicating its robustness and accuracy in predicting daily ride volumes. The decision tree regression, our next best model, identified maximum temperature, total precipitation, and day length as the most important features.

Recommendations:

Key performance metrics indicate that our model achieves high accuracy, providing actionable insights for Mobi's operations team. Future efforts should focus on refining the model and exploring additional data sources to further enhance prediction accuracy and operational efficiency.

Our analysis revealed that all models performed significantly better after detrending the data, indicating that the trend removal method can be substantially improved. We recommend adopting a more granular approach to predict usage per bike station and at different times of the day, which could facilitate the implementation of dynamic pricing strategies.

Although our current model is tailored to Mobi, we believe there are general trends in bikeshare usage applicable across companies and cities. Incorporating data from multiple sources would enable the development of more accurate and robust models. Expanding the dataset to include diverse geographic and operational contexts will provide broader insights and enhance the predictive power of the model.

Acknowledgments:

We sincerely thank the Erdős Institute and our advisor, Dyas Utomo, for their assistance with this project.