MIXED APPROACH FOR BIKE AVAILABILITY FORECAST

Ignacio Sevillano Muñoz
Atos Research and Innovation
Madrid, Spain

ABSTRACT

We propose a predictive analysis model for bike-sharing systems based on a mixed model of statistical and machine-learning approaches. The model assumes a Poisson distribution on bike station arrival and departures. The model is validated by comparing the predicted arrival and departure rates to the actual rates. The case study results show that our model can outperform other naïve approaches using the mean on arrival and departure rates.

KEYWORDS

Bike Availability Forecast, Predictive Analysis, XGBoost

1. INTRODUCTION

The advancements in IoT and AI have brought about a significant transformation in our interaction with the physical world. It has enabled businesses to optimize their processes and make well-informed decisions. In this regard, the GreenMov initiative deserves accolades for its efforts toward developing eco-friendly transportation services that enhance the standard of living in urban areas. This article sheds light on the bike-station availability forecast (BAF for short) service that employs predictive analytics to ensure that bikes are readily available for users in Flanders (Belgium) and Nice (France).

Bike-sharing systems are an effective way to reduce traffic congestion and carbon emissions in urban areas. They work by providing a network of bike stations $\{s_i\}_{i=1}^{n}$ where users can rent and return bikes. Each station s_i has a capacity c_i, which is the maximum number of bikes that can be parked there. We represent the number of bikes at a certain station s_i at a given time t, or availability, by $a_i(t)$. Furthermore, the forecast we refer to in this context is an estimation of the station availability denoted as $\widehat{a_i}(t)$.

In this article, we introduce a novel predictive availability service designed to tackle the issue of bike availability forecast in bike-sharing systems. Our proposed solution combines statistical foundations within a machine learning (ML) framework. By leveraging the inherent capabilities of Decision Trees algorithms, our approach enables us to provide accurate predictions while also accommodating the inclusion of additional regressors in the future. This article opens up new possibilities for the development of user-centric and data-driven services that can make our cities more livable and sustainable., and contribution of the paper.

This article begins (Section 2) with the introduction of a statistical model for bike availability in bike-sharing systems and its validation and forecast processes. The subsequent section (Section 3) presents and evaluates the case study of Blue-bike, in Belgium. Finally, the article concludes by proposing improvements and further applications of this service.

Gast, & Bierlaire, 2014, proposed a probabilistic forecast of bike-sharing station availability using a queuing-theoretic approach. Zhu, & Rasouli, 2016, also used a queuing-theoretic approach to design bike-sharing systems. Wang, W. & Madanat, S., 2017, developed a stochastic optimization model for bike-sharing system design.

Chen, & Guestrin, 2016, developed XGBoost, a scalable tree-boosting system. Guo, Chen, He, Ma, & Liu, 2017, studied the importance of tree diversity in boosting. Sun, & Srivastava, 2017, presented a comprehensive survey on Extreme Gradient Boosting.

For other approaches, (Lin, & Dou, 2016) proposed a research on urban public bicycle demand forecast based on a network perspective. Patil,, Musale, & Rao, 2015, used RandomForests to predict bike share demand. Duan, & Wang, 2017, proposed a moment-based availability prediction for bike-sharing systems.

2. STATIONS AS WAITING QUEUES

2.1 Regression Method

In this subsection, we present the statistical model for bike stations, which is modeled as a Markovian queuing system. Rather than focusing on the station's availability $\boldsymbol{a_i(t)}$, we prioritize the arrival and departure rates, $\boldsymbol{\alpha_i(t)}$ and $\boldsymbol{\delta_i(t)}$ respectively, and define them as piecewise functions composed by 30-minute time intervals, i.e. $\boldsymbol{24\,[hours]\cdot 2\frac{[intervals]}{[hour]}=48}$ intervals. Thus, we need to determine $\boldsymbol{\alpha_{i,j}(t)}$ and $\boldsymbol{\delta_{i,j}(t)}$, with $\boldsymbol{i=1,2,3,\dots,n}$ and $\boldsymbol{j=1,2,3,\dots,48}$. We assume that the arrival and departure rates of bikes at any station follow a Poisson process, which is characterized by the Poisson Probability Density Function:

$$f(k;\lambda)=\frac{e^{-\lambda}\lambda^{k}}{k!}$$

To verify if the arrival rate $\alpha_{i,j}(t)$ conforms to a Poisson distribution, we employ the Kolmogorov-Smirnov (K-S) test. Firstly, we generate an empirical probability function $g_{\alpha_{i,j}}(k)$ for each time interval. Then, using the previous equation for the Poisson pdf. with $\lambda_{\alpha_{i,j}}$ denoting the mean of arrivals at interval j in station s_i, we compute the cumulative distribution functions $G_{\alpha_{i,j}}(k)$ and $F\left(k;\lambda_{\alpha_{i,j}}\right)$. Subsequently, we calculate the K-S test statistic, defined by

$$D:=\sup_{k}\left|F\left(k;\lambda_{\alpha_{i,j}}\right)-G_{\alpha_{i,j}}(k)\right|$$

We can then compare the test statistic D with the critical value $D_{0.95}$ at a chosen significance level 0.95. If D exceeds $D_{0.95}$, we reject the null hypothesis that the observed data follows a Poisson distribution. Same procedure applies for departures $\delta_{i,j}(t)$.

2.2 Regression Method

In Section 3, we will demonstrate that the majority of stations exhibit arrival and departure rates that can be approximated by a Poisson distribution during the daytime with fixed λ as the mean of arrivals or departures. However, our objective is to enhance this naive parameter by employing a machine learning model, thereby establishing a more flexible framework that incorporates various regressors. Specifically, our regression approach relies on the utilization of *XGBoost* decision trees, where lag values are chosen based on correlation analysis.

In our analysis, we discovered that in addition to the typical lag variables, such as the previous day, previous week, or previous hours, there exists a significant correlation between the departure rates $\delta_{i,j}(t)$ and the arrival rates $\alpha_{i,j}(t)$ in the time interval 6 to 8 hours later. This observation can be attributed to individuals using bicycles for their commute to work and relying on the bike-sharing service for their return journey home as well.

To estimate the expected station availability, we generate piecewise functions for the predicted arrival and departure rates, denoted as $\widehat{\alpha_i}(t)$ and $\widehat{\delta_i}(t)$ respectively, by utilizing the forecasted values from each time interval. Subsequently, we perform an integration of these functions over time to calculate the anticipated availability with the desired level of precision. Then, for each station s_i

$$\widehat{a_i}(t) \coloneqq a_i(t_0) + \int_{t_0}^{t} \widehat{\alpha_i}(x)dx - \int_{t_0}^{t} \widehat{\delta_i}(x)dx$$

The estimation for the availability $\widehat{a_i}(t)$ at time t is defined for any $t \in R$, where t_0 represents the last available data time. This allows us to tailor the forecast resolution to specific needs. Furthermore, this flexibility is advantageous in cases where bike stations exhibit low activity levels, as it enables us to consider wider time intervals and assume a Poisson distribution underlying the data.

3. RESULTS AND PERFORMANCE

This section showcases the outcomes of our predictive analysis model for bike-sharing systems and offers insights into their implications for system design and operation. We evaluate the performance of our model through a case study: Blue-bike, utilizing data from Belgium.

We commence by validating the statistical model and subsequently present our results utilizing the coefficient of determination (R^2). This metric serves as an indicator of the accuracy of our forecast service, providing short-term precision assessments.

3.1 Blue-Bike

Blue-bike is a national bike share scheme in Belgium founded in 2011. It operates bike-sharing services through over 71 locations across the country. In 2019, Blue-bike recorded 276,000 rides with 20,000 members. Surveys indicate that 93% of members combine Blue-bike with train journeys, and 32% of rides were previously unmade or replaced car trips.

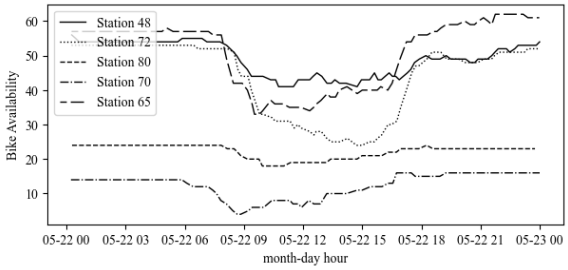

Figure 1. Bike availability at 22nd of May

For our analysis, we utilize one year of available data. Initially, we validate our model by distinguishing between workdays and non-working days. To provide clarity, we present our Kolmogorov-Smirnov (KS) tests in a 3D plot, where the x-axis represents the hour of the day, and the y-axis represents the p-value. We generate separate plots for arrival and departure rates, considering both workdays and non-working days, resulting in four distinct figures.

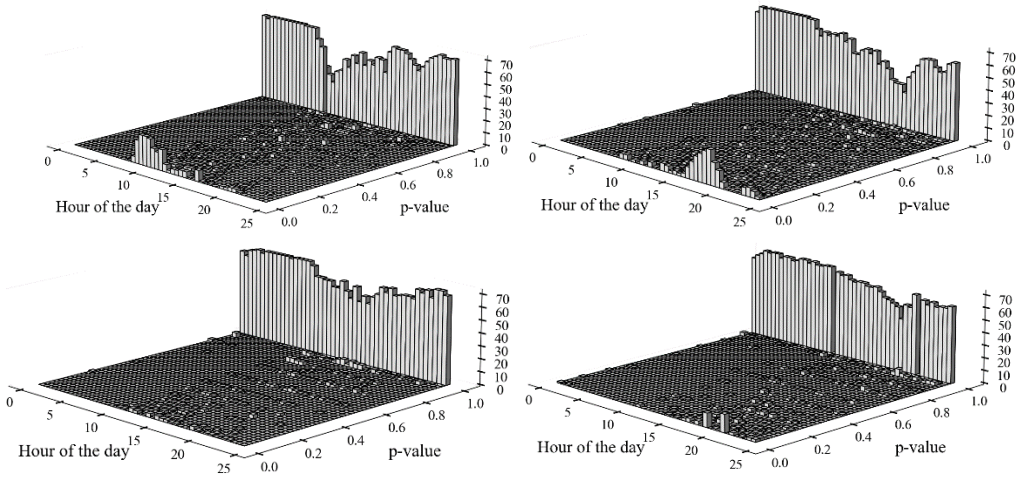

Figure 2. KS-tests for arrival and departure rates. From left to right, from top to bottom: Departure rates - Workdays, Arrival rates - Workdays, Departure rates - Non-working days, Arrival rates - Non-working days

In the given figures, we can observe that the arrival rates during the afternoon and the departure rates in the morning (Figures 2a} and 2b), which represent the busiest time periods overall (refer to Figure 1), do not seem to conform to a single Poisson distribution with a fixed mean parameter, as initially suggested in Section 2. To address this issue, employing Machine Learning models would offer a valuable opportunity to enhance the accuracy of these rates.

To conclude this section, we present the results obtained through training on multiple datasets (Figure 3). We compare two sets of results: the first one (left) represents the naive approximation, which yields an average R^2 value of 0.95, while the XGBoost approximation (the method showcased in this article, on the right) achieves an average R^2 value of 0.96. This improvement may seem subtle at first glance, but it is more significant than it appears. It should be noted that most of the test data consists of moments during the day when the arrival and departure ratios are trivial, such as nighttime periods. Hence, achieving higher accuracy in such cases demonstrates the robustness and effectiveness of the proposed XGBoost method.

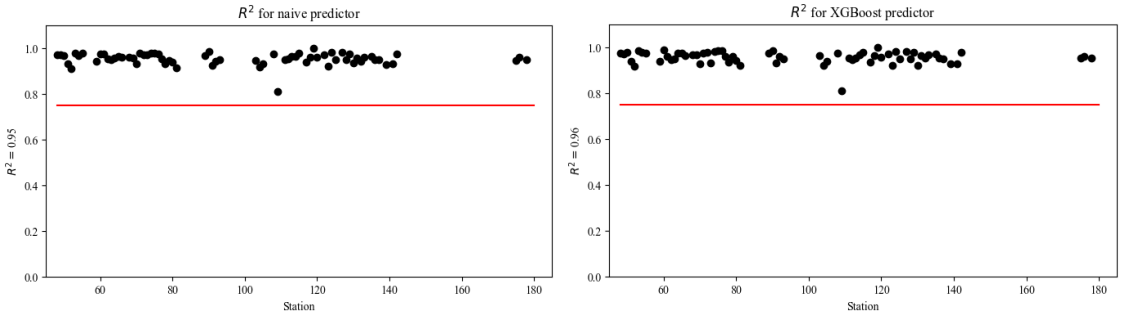

Figure 3. Results. Left: Naive approach. Right: XGBoost approach

4. CONCLUSION

This article introduces a novel predictive availability service for bike-sharing systems, employing statistical modeling and machine learning techniques. The proposed solution demonstrates high accuracy in predicting bike availability, with a strong correlation between the predicted and actual values. The approach presented here opens up possibilities for developing user-centric and data-driven services, contributing to the creation of more livable and sustainable cities.

Moving forward, further improvements and research can be conducted to enhance the predictive model. One potential area of exploration is studying the long-term behavior of bike availability predictions. Also, considering factors such as seasonal variations, events, and holidays.

Additionally, investigating the application of the predictive service in fleet redistribution strategies could optimize bike-sharing operations and improve user experience. Furthermore, extending the service to include other aspects, such as real-time demand forecasting and route optimization, could provide valuable insights and additional functionality for bike-sharing systems.

REFERENCES

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

Duan, Z. & Wang, L. (2017). Moment-based availability prediction for bike-sharing systems. *IEEE Transactions on Intelligent Transportation Systems, 18(11)*, 2832-2841.

Gast, M. & Bierlaire, M. (2014). Probabilistic forecast of bike sharing stations availability. *Transportation Research Part B: Methodological, 68, 192-207.*

Guo, H., Chen, T., He, T., Ma, Y., & Liu, Y., (2017). On the importance of tree diversity in boosting.

Lin, Y. & Dou, W. (2016). Research on urban public bicycle demand forecast based on network perspective. *Journal of Intelligent Transportation Systems, 20(3)*, 234-245.

Patil, A., Musale, K. & Rao, B. (2015). Bike share demand prediction using RandomForests. *International Journal of Innovative Science, Engineering and Technology, 2(2)*, 83-88.

Sun, Y. & Srivastava, A. (2017). A comprehensive survey on Extreme Gradient Boosting.

Wang, W. & Madanat, S. (2017). A stochastic optimization model for bike sharing system design. *Transportation Research Part B: Methodological*, 105, 116-134.

Zhu, J. & Rasouli, A. (2016). A queuing-theoretic approach for bike sharing system design. *Transportation Research Part B: Methodological, 84*, 16-31.