International Journal of Business Analytics
Volume 9 • Issue 1

Prediction of Bike Share Demand by Machine Learning:
Role of Vehicle Accident as the New Feature

Tae You Kim, Catholic University of Korea, South Korea
Min Jae Park, Ajou University, South Korea
Jiho Shin, Business School Lausanne, South Korea
Sungwon Oh, Business School Lausanne, South Korea

ABSTRACT

In the fourth industrial revolution period, multinational companies, and start-ups have applied a sharing economy concept to their business and have attempted to better serve customer demand by integrating demand prediction results into their business operations. For survival amongst today's fierce competition, companies need to upgrade their prediction model to better predict customer demand in a more accurate manner. This study explores a new feature for bike share demand prediction models that resulted in an improved RMSLE score. By applying this new feature, the number of daily vehicle accidents reported in the Washington, DC area to the random forest, XGBoost, and LightGBM models, the RMSLE score results improved. Many previous studies have primarily focused on feature engineering and regression techniques within a given dataset. However, this study is meaningful because it focuses more on finding a new feature from an external data source.

KEYWORDS

Bike Share Demand Prediction, Daily Crashes Reported in Washington, DC, Machine Learning, New Feature Exploration, Sharing Economy

1. INTRODUCTION

Predicting customer demand accurately in the data-driven business environment of the mobility industry is a key factor of success (Sohrabi et al., 2020; Wessel, 2020). In the data technology period, it is very easy to find multinational companies that provide their various services to customers based on demand prediction results. In the short term, companies have achieved success because they have sufficiently predicted demand based on internal and external features. Digital transformation is occurring in most industries around the world based on machine learning and deep learning algorithms (Veres & Moussa, 2019; Wang et al., 2019). In particular, rapidly growing start-ups are often making their important business decisions based on data and algorithms, not on management experience or intuition.

DOI: 10.4018/IJBAN.288513

In a modern society where environmental problems are increasingly serious, the bike is an eco-friendly means of transportation that benefits from individuals maximizing its exercise effects during the COVID-19 epidemic. The bike has attracted worldwide attention since a few years ago as it has been transformed into a "shared" good. It is not just for leisure purposes, but for the transportation itself (Böcker et al., 2020; Gu et al., 2019). Shared bike programs in many cities around the world are considered a means of responding to climate change and energy crises when implementing sustainable transportation systems. A shared bike system is an important and necessary means to promote the use of bikes and further implement an efficient urban transportation system. However, despite its many advantages, the number of cities and businesses that are reluctant to introduce a shared bike program is increasing due to the disadvantages of high fixed costs such as installation cost or operating cost. Thus, an accurate demand prediction is required for the continuous operation of a bike sharing program. One of the earliest shared economy models in the mobility industry was the bike sharing system. In particular, Capital Bikeshare, which launched the bike rental business in the Washington, D.C. area, hosted a data science competition in the Kaggle Competition platform to predict customer bike share demand (Fanaee-T & Gama, 2014). In response, a number of efforts have been made by data scientists and data analysts around the world to predict demand by using various data mining techniques.

Shared bikes are emerging as an alternative means of transportation that can comprehensively improve environmental problems, traffic jams, poor quality of life caused by car-oriented transportation systems, and can guarantee practical effects (de Chardon, 2019; Gu et al., 2019). In addition, shared bikes have several advantages when establishing an integrated public transportation system in connection with public buses and urban rail systems. First, the average distance of a bike trip is around 3-5 km (last mile), but it is possible to travel longer distances in connection with public transportation. Secondly, the catchment area of public transportation can be extended from an existing pedestrian area to an area coverable by bike. Third, during non-adjacent hours (except rush hour when traffic demand is concentrated during the day), highly demanded shared bikes can accommodate leisure traffic on existing transit, which can increase the overall utilization rate of public transportation. Bikes are the best means to realize an integrated public transportation system by supporting traffic in combination with walking or public transportation (Böcker et al., 2020; Gu et al., 2019). An integrated public transportation system can increase the availability of public transportation by both merging public transportation services, such as buses or metros, into high capacity corridors where economies of scale are built and strengthening cooperation with corridors or feeders using bikes (Gu et al., 2019). Therefore, it is imperative to predict the demand for shared bikes to efficiently implement an integrated public transportation system that includes shared bikes.

In previous studies, many focused on the importance of features within the models, such as the distance between rental and return location (Ma et al., 2020; Younes et al., 2020). These studies primarily focused on engineering feature techniques and statistical modeling within given datasets. Therefore, there were model limitations from previous studies since they only used given datasets. This study, however, focuses on exploring a new feature from an external data source. Existing studies have predicted demand for bikes through statistical methodologies, but few have applied machine learning methodologies, which have recently been in the spotlight. The purpose of this study is to present a machine learning prediction model that uses variables that affect the demand for shared bikes. Furthermore, from the perspective of an integrated public transportation system, a predictive model that includes these variables is presented with the belief that traffic accident occurrence affects the demand for shared bikes as an alternative transportation system. The analysis results are expected to provide policy implications for efficient operation of a city's integrated public transportation system.

The sequence of study is as follows. Section 2 provides a review of literature and reviews the history of bike share programs across the globe and in the U.S. and the prediction of bike share demand by various data mining and statistical approaches. Section 3 covers methodology and discusses data set overview, exploratory data analysis, and data preprocessing. In addition, the new feature is introduced