INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Predicting Bicycle Arrivals in a Bicycle Sharing System Network: a Data Science Driven Approach Grounded in Zero-Inflated Regression

Weidmam Milagres Leles

Master's in data science,

Supervisor:
Ph.D., Ana Maria de Almeida, Associate Professor,
ISCTE - University Institute of Lisbon

Co-supervisor:
Ph.D., Diana Aldea Mendes, Associate Professor,
ISCTE - University Institute of Lisbon

October, 2022

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Predicting Bicycle Arrivals in a Bicycle Sharing System Network: a Data Science Driven Approach Grounded in Zero-Inflated Regression

Weidmam Milagres Leles

Master's in data science,

Supervisor:
Ph.D., Ana Maria de Almeida, Associate Professor,
ISCTE - University Institute of Lisbon

Co-supervisor:
Ph.D., Diana Aldea Mendes, Associate Professor,
ISCTE - University Institute of Lisbon

October, 2022

To my much-loved parents, who are the reasons of my life and who teach me the true meaning of unconditional love; to my precious brothers, with whom I share love and closeness enough to tell each other what we need to hear and not what we would like to hear; to my maternal grandfather, who I had the pleasure of enjoying part of my life with; to my maternal grandmother, who I admire so much for her good heart; and to my paternal grandparents, whom I did not meet in life, but of whom I am very proud of for their life story that inspires me so much

Acknowledgment

First and foremost, I would like to acknowledge my supervisor, Professor Ana Maria de Almeida, and my co-supervisor, Professor Diana Aldea Mendes. Their dedication, availability and keen knowledge was undeniably essential to the successful completion of this study.

My acknowledgement also goes to the incredible professors who have accompanied me throughout these two years, from whom I have had the opportunity and pleasure to learn much of what I know today.

I would also like to acknowledge my professional colleagues and, in particular, my head manager, who was very supportive and made it possible for me to align my professional duties in a period of intense academic commitment.

I also would like to express my gratitude to my family, especially to my parents and brothers, for all the emotional support, encouragement, and for facing my goals as if they were their own. Without their motivation, none of this would have come true, they are the ones that drive each of my achievements.

To my childhood friends, to the friends I have met throughout my life, scattered around the world, and to my friends in Lisbon, who give me strength for the challenges of everyday life, my sincere thanks. Each one of them contributes in diverse ways, but in an incredibly special manner, towards the achievement of this goal.

To my classmates, whom I have enjoyed getting to know and working together with during these two years of the master's program, I also want to thank for all the mutual support.

Conciliating academic aspirations with professional life, with the social roles to be fulfilled, living a healthy life, and still having time for leisure is not an easy task at all, so I extend this acknowledgement to everyone who contributed, in one way or another, for this path to end successfully.

Resumo

A adoção de sistemas de bicicletas partilhadas (BSS) vem crescendo com o objetivo de melhorar a forma como as pessoas se deslocam pelas cidades, mas também para estimular o desenvolvimento de uma mobilidade urbana mais sustentável. Para o bom funcionamento de um BSS é importante que haja bicicletas permanentemente disponíveis nas estações para os utilizadores iniciarem as suas viagens, pelo que a literatura tem empreendido esforços, sob a ótica do operador do serviço, para melhorar o processo de redistribuição das bicicletas e assim garantir a sua disponibilidade nas diferentes estações. Como a garantia de bicicletas disponíveis não pode ser assegurada, este trabalho propõe-se desenvolver, sob a ótica do ciclista, uma prova de conceito sobre a viabilidade de informar o utilizador acerca da possibilidade de iniciar uma viagem num intervalo de tempo pré-definido. As principais contribuições deste trabalho são: (i) a capacidade de previsão de quantas bicicletas chegarão a uma determinada estação é uma melhoria viável para os BSS, (ii) os modelos desenvolvidos através da aproximação Zero-Inflated Regression são um caminho que pode ser explorado para melhorar a previsão e (iii) contributo metodológico inédito à literatura sobre os BSS com foco no poder decisório do utilizador final sobre se será, ou não, possível iniciar uma viagem em breve.

Palavras-chave: Sistemas de Bicicletas Partilhadas, Zero-Inflated Regression, Séries Temporais

Abstract

The adoption of bicycle sharing systems (BSS) is growing in order to improve the way people move around cities, but also to stimulate the development of a more sustainable urban mobility. For the proper functioning of a BSS, it is important to have bicycles permanently available at the stations for users to start their trips, so the literature has undertaken efforts, from the perspective of the service operator, to improve the process of redistribution of bicycles and thus ensure their availability at the different stations. Since the guarantee of available bicycles cannot be assured, this work proposes to develop, from the cyclist's perspective, a proof of concept on the feasibility of informing the user about the possibility of starting a trip in a pre-defined time interval. The main contributions of this work are: (i) the ability to predict how many bicycles will arrive at a given station is a feasible improvement for BSS, (ii) the models developed through the Zero-Inflated Regression approach are a path that can be explored to improve prediction and (iii) unprecedented methodological contribution to the literature on BSS focusing on the end-user's decision power about whether or not it will soon be possible to start a trip.

Keywords: Bicycle Sharing Systems, Zero-Inflated Regression, Time series

Index

Index Glossary of Acronyms

AIC	Akaike Information Criterion
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Auto Regressive Integrated Moving Average with exogenous variables
BSS	Bicycle Sharing Systems
CRISP-DM	Cross Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
FP-Growth	Frequent Pattern-Growth
GBM	Gradient Boosting Machine
GEV	Generalized Extreme Value
GPC	Gaussian Process Classification
GPS	Global Positioning System
IQR	Interquartile Range Rule
LSTM	Long short-term memory
MAE	Mean Absolute Error
PEST	Political, Economic, Socio-cultural and Technological
R-GCN	Relational Graph Convolutional Networks
RMSE	Root-mean-square error
SVM	Support Vector Machines
VIF	Variance Inflation Factor
ZIR	Zero-Inflated Regression

Introduction

The adoption of bicycle sharing systems (BSS) has been growing, improving the way city residents travel around town, and stimulating the development of sustainable urban mobility. The bike sharing system allows users to rent a bike at one of the bike docks found in the city, use it for their trips and return it to any bike dock (Macioszek et al., 2020). It is important to note that there are also BSS that are dockless, i.e., the beginning and end of the journey are not associated with a fixed bike dock, thus providing more freedom to the cyclist to decide the point of departure and arrival (Liu & Pelechrinis, 2021). This work focuses only on fixed dock BSS.

The transition to multimodal and shared transport has been seen by urban planners and transport system engineers as an important way to drive cities to thrive and pave the way for a sustainable future (Liu & Pelechrinis, 2021). Bike Sharing Systems have been adopted in many cities as an alternative to other manners of transport, whether public or private (Cenni et al., 2021). According to Ge et al. (2020) bike-sharing systems are short-term rental services that provide convenience to users and serve as an effective means of solving significant urban mobility challenges such as the "last mile" problem[1], connection to public transport and reduction of dependence on a private car in some situations.

Furthermore, as Ge et al. (2020) point out, using bicycle-sharing systems also contributes to keeping users' health and reducing environmental pollution. Notwithstanding the benefits of the BSS mentioned above, these systems bring some challenges that restrain users from fully enjoying their benefits. Such as the irregular distribution of bicycles at stations and the difficulty in anticipating whether bicycles will be available to start a journey or if there will be free docks to return them to the destination bike rack (Cenni et al., 2021).

Forecasting demand and availability become a key element for the proper functioning of BSS, and several authors made efforts to make accurate forecasts of the demand, availability, and redistribution of bicycles (Almannaa et al., 2020; Ashqar et al., 2017; Cavallaro & Tramontana, 2021; Cenni et al., 2021; Z. Chen et al., 2021; Feng et al., 2017; Liu et al., 2019; Ruffieux et al., 2019). Even so, Sardinha et al. (2021) point out that recent contributions have shown limited success in forecasting demand for this type of transport, as spatial,

[1] Additional short distance to destination that one must travel after leaving a public transport service (Ge et al., 2020)

meteorological, situational, and seasonal contexts play essential roles when it comes to forecasting demand and availability of shared bicycles.

Although forecasting the demand for bicycles, precisely to help ensure the availability of bicycles at bike docks, has received significant attention in the literature, many of these problems are dealt with at a more operational level, with the aim of helping operators to improve their service provision. This work intends to study the availability of bicycles from the user's perspective and bridge this gap in the literature, that is, since the availability of a bicycle to start a journey is not guaranteed, it would be important to supply the user with information about the possibility of starting a journey shortly. In other words, providing the user with information about whether a bicycle will arrive at that particular station in the next minutes, so that with the arrival of this bicycle, it will be possible to start a journey.

The goal of this work is to develop, from the cyclist's perspective, a proof of concept on the feasibility of informing the user about the possibility of starting a trip in a pre-defined time interval by predicting whether bicycles will arrive at a given station. Along these lines, it is intended to predict, for each station, whether a bicycle will arrive soon. If the forecast is positive, that is, if the model predicts that bicycles will arrive in in a while, it is also important to approximate how many bicycles will arrive. A specific goal of this study is to verify whether models based on a zero-inflated probability distribution, called zero-inflated regression (ZIR) model, are suitable for this type of problem, as the bicycle count is non-negative and has a high proportion of zeros. Thus, the problem is twofold: first a classification task to deal with the high excess of zeros and later a regression problem to handle non-zero values.

This study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) process model, which is divided into six steps: (i) Domain understanding, which is expressed here both by the framing of the theme in this Introduction, and through the reflection of the literature review concerning to the problems inherent to the BSS and the use of data science methodologies to solve these problems; (ii) Data Understanding, which consists of the interpretation and analysis of the datasets to better relate it with the goal of this study; (iii) Data preparation, a step that consists of transforming, cleaning, aggregating, and reformatting the original data for exploration and modeling purposes; (iv) Modeling phase, in which prediction models are built to train and test model selection and development; (v) Evaluation stage, in which the evaluation of the models is done in order to assess if they meet the goals of this study. That is, if the models can predict how many bicycles will arrive in the next minutes and if the models integrated into the Zero-Inflated Model framework are more capable of performing such a prediction; (vi) Deployment, as the last stage of CRISP-DM, it was intended to trace the

implementation path aimed at improving the level of service delivery of bike-sharing systems at the user level.

Having completed all the CRISP-DM stages, conclusions were drawn, and the main contributions and limitations of the work were aligned considering the aim of the feasibility of informing the user about the possibility of starting a trip in a shortly and on the contributions that the models based on ZIR can bring to this matter.

CHAPTER 2

Systematic Review of Literature

2.1. Research and Selection Criteria

This section aims to clarify the methodology used to conduct the Systematic Literature Review. The starting point has been that of understanding whether it is possible, through zero-inflated data science techniques, to predict the arrival of a bike to a BSS station.

To conduct the Systematic Review of the Literature to promote a thorough, objective, and reproducible research, it was chosen, among the various sources that can be consulted for a systematic review, the SCOPUS bibliographic database, as it indexes many scientific journals and can be easily accessed.

The sample of papers composing the literature review conducted in this work originated from the junction of four queries, as can be seen in Table 1. Each query has a distinct objective and selection criteria which together form the content needed to understand how the prediction of bike arrivals in each station, fits into the current scientific context.

Table 1 - Used queries and their purposes

Query purpose	Query	# Documents analyzed	# Selected documents	Selection criteria
Query 1 – Obtain literature on forecasting the availability of bike-sharing systems	TITLE-ABS-KEY(bike OR bicycle-sharing OR "bike-sharing" OR "Bicycle-sharing system" OR "Bike-sharing system" OR "Bike sharing" OR public W/2 bike AND prediction OR predicting OR predict AND availability)	49	19	Title or abstract that directly addresses the issue of predicting bike availability
Query 3 – Obtain literature on waiting time to pick up a bike	(ALL (prediction OR predict OR predicting AND "waiting time" OR waiting OR wait AND bike OR bicycle)) AND AUTHKEY ("Bike sharing demand" OR "Bike-sharing systems" OR ~waiting)	20	3	Title or abstract that directly mentions the issue of the user's waiting time to start a trip
Query 4 – Obtain literature on context awareness applied to BSS	ALL ("context-aware" OR "spatiotemporal" OR "situational context" OR "contextual factors" AND bike OR "bike-sharing") AND AUTHKEY ("Data Science") AND (LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2021))	10	2	Title or abstract that directly mentioned the issue of situational context applied to BSS
Query 5 – Obtain literature on Zero-Inflated Model	TITLE-ABS-KEY("~Zero-Inflated " AND bike OR bicycle OR bss OR "~Bike Share Systems")	17	4	Title or abstract that directly addresses the use of the Zero-Inflated model in relation to the prediction of use of BSS

The SCOPUS database is used to search for relevant material published from the last 10 years, considering keywords related to bike availability prediction in bike-sharing systems. Except for query number four, for which the research was limited only to publications from the years 2021 and 2022, since the intention was to collect the most recent practices in the use of external variables.

All queries' results were analyzed by reading their titles and abstracts to select the documents that were in line with the aim of this study and with the selection criteria of each one of the queries, as depicted in Table 1. Twenty-eight documents were chosen to be kept in Mendeley, a reference manager software, for further reading of the abstract, introduction, results and conclusions of each of the selected documents.

2.2. Discussion

There are several studies dedicated to forecasting bicycle demand in bike sharing systems, i.e. the expected number of bicycles to be collected and returned to a BSS station. (L. Chen et al., 2016; E et al., 2020; Liu & Pelechrinis, 2021; Lucas & Andrade, 2021). Bacciu et al. 2017 pay attention to the time a user will need to wait to get a bicycle or return it. In fact, for other public transport systems, such as trains, subways, and even buses, it is common to have a panel showing the waiting time for the next transportation service to arrive.

In this regard, scholars have highlighted that when the bike start station is empty, that is, there is no way to start a journey as no bike is available, the user experience can be improved by informing the users about (i) how the situation will evolve, (ii) whether a bicycle will arrive, and (iii) how long it will take to arrive (Bacciu et al., 2017; Soheil et al., 2020; J. Wang et al., 2022). As highlighted by some authors (J. Wang et al., 2022), the BSS literature has not paid much attention to the time an end user has to wait at a given station. The lack of efficiency in transport systems often arises from uncertainty, such as the lack of a bicycle to start a journey in a BSS (B. Chen et al., 2013). In order to correct this inefficiency, researchers explicitly included, in their predictive models, exogenous variables, such as weather and traffic information, which significantly helped to predict the waiting time till a bike trip could be started or till a bicycle could be left at an available bike rack at the end of a trip (B. Chen et al., 2013). Although the focus may not precisely be the waiting time till a bike trip starts, other authors have also equated the problems studied with end-users, in addition to system operators,

regarding increasing the predictability of the availability of resources for the proper functioning of bike-sharing systems (Gast et al., 2015; Salih-Elamin & Al-Deek, 2020).

Forecasting the availability of bicycles and free docks in bike racks has been central to ensuring the success of such bicycle-sharing systems (Demidova et al., 2022). The demand for bicycles and bicycle racks is an unbalanced factor during the day, with significant pressure in some stations.

What is currently being introduced as a novelty is the use of data that incorporates context awareness to make the most accurate demand forecasts (Demidova et al., 2022). The incorporation of context awareness is critical for more reliable predictions (Sardinha et al., 2021). For these authors, the contributions in this area have seen limited success in predicting the demand for the bike-sharing service because they do not include exogenous explanatory variables in the model. There is a strong dependence between bicycle demand and the meteorological context. For example, in addition to the situational aspect of the seasons and the inability of most predictors to consider the effects of high or low demand on nearby stations, it has a direct effect on the demand forecast and, consequently, on the availability in each bike rack (Sardinha et al., 2021).

Also with the goal of providing insights into the rebalancing operations of the bicycle network, developed a statistical model applied to the bicycle system of the city of Lisbon, in Portugal, based on the Zero-Inflated Model and Hurdle model, to predict the number of trips that will occur between two stations at a certain time, as well as the duration of these trips (Lucas & Andrade, 2021). Furthermore, part of the success of the model developed by Lucas & Andrade (2021) is due to the incorporation of contextual variables in the model: explanatory variables, including the weather, effects of intramodality in transport, the effects derived from the location of the stations due to their proximity to different types of points of interest, and the effective travel time (Lucas & Andrade, 2021).

Given the importance of these external factors, some authors developed a new method on how to acquire, consolidate and incorporate different variables for the context-enriched analysis of traffic data, as many different situational contexts have been considered in previous research works to analyze the dynamics of traffic (Cerqueira et al., 2021). This data includes weather information, potential occurrences from Twitter data, accident records, sports matches, festivals, and other events inferred from social data (Cerqueira et al., 2021). Therefore, analyzing data from trips together with context data brings significant benefits in the development of more robust models capable of predicting the demand and availability of bike-sharing systems with greater accuracy. That is, traffic sensors, global positioning system (GPS)

trajectories and location-based social media data provide several sources of information that help to detect and analyze spatio-temporal events. (Lam et al., 2019). The heterogeneous combination of GPS data with geotagged tweets produced satisfactory results for inferring human mobility, especially during unprecedented events, according to a study carried out by Miyazawa et al. (2019). However, it is important to note that data from social media are subject to age bias and inconsistencies in data collection (Thu et al., 2017).

Corroborating that demand for bike-share is affected not only by common contextual factors such as time of day and weather, but also by opportunistic contextual factors such as social events, for example, Chen et al., (2016) proposed a dynamic cluster-based structure to forecast excess demand. Depending on the context, they developed a weighted correlation network to model the relationship between bicycle stations and dynamically group neighboring stations with similar bicycle usage patterns into clusters. Then, they adopted the Monte Carlo simulation to predict the probability of excessive demand for each cluster.

Most of the forecasting models developed on bike-sharing systems, as highlighted by Zhang et al. (2018), are based on their own usage history, which impairs prediction when some sudden event occurs. In this sense, Zhang et al. (2018), when considering these systems as an integral part of the public transport system, use the interrelationship between bike sharing and other transport to capture the impact of sudden events, thus improving the predictive capacity of the developed models. Complementarily and to corroborate with this vision, it is also important to mention that, as Liu & Pelechrinis (2021) emphasize, most studies on demand forecasting in bike-sharing systems consider only observed demand, that is, the demand reflected in the trip data recorded by the system. However, total demand also includes trips that never took place but were intended by users. In this way, Liu & Pelechrinis (2021), developed a regression model capable of predicting the difference between the total demand and supply of bicycles.

Some authors used LSTM (Long short-term memory) models in their studies. Liu et al. (2019) highlighted that LSTM models have the advantage of being able to be built with multiple input data and with multi-time steps, which improves prediction accuracy. Jiang et al. (2019) point out that bicycle trips can be formulated as a time series and, therefore, LSTMs have been used successfully to train complex time series data in a variety of applications, such as road traffic forecasting, for example. LSTM models are particularly suitable for classification, processing and forecasting tasks with time series data that have a time gap between notable events of unknown size or duration, just as in bike-sharing problems (Jiang et al., 2019). Zhang et al. (2018) also developed a prediction model based on neural networks with the LSTM

architecture. According to these authors LSTM is adopted because it shows excellent performance for learning long-term temporal existing dependencies on BSS (Zhang et al., 2018).

Salih-Elamin & Al-Deek (2020) when predicting short-time and short-distance BSS trips duration throughout the day under weather conditions concluded that the Stepwise Multiple Linear Regression is the model that has the best performance compared to Autoregressive Integrated Moving Average (ARIMA) and Auto Regressive Integrated Moving Average with exogenous variables (ARIMAX). Cenni et al. (2021) with a similar goal but a focus on long-term bike availability predictions, realized that models based on Gradient Boosting Machine (GBM) offer a solid approach to implementing reliable predictions in conditions where there are few bikes available. As well as (Sathishkumar et al., 2020) when comparing five[2] different modeling techniques, with the objective of guaranteeing availability and accessibility of bicycles in bicycle racks, concluded that the GBM, even with different combinations of predictors, is the best model they developed.

Some authors have decided to apply deep learning to develop algorithms based on relational graph convolutional networks (R-GCN) to reposition bikes and manage bike-sharing systems (Yoshida et al., 2019). The proposed algorithm consists of three parts: predict the check-outs and check-ins of bicycles; find the ideal number of bicycles to satisfy the entire demand of the system; determine the truck route to rebalance bikes in the system. The results suggest that the developed model can propose routes that are more realistic to be implemented compared with the current models (Yoshida et al., 2019). When trying to improve the routes of bicycle rebalancing trucks, other authors also used neural networks; but they assessed a Multilayer Perceptron algorithm, which was also able to perform the prediction with greater accuracy (Ruffieux et al., 2019).

Authors such as Cavallaro & Tramontana (2021), in the context of user assistance software, propose an approach using frequent pattern-growth (FP-Growth) to estimate where bicycles will be shortly, leveraging the use of such transport modality, since users are informed in advance about the availability of bicycles. Ensuring that a user arrives at a station has also received attention from other authors, who have developed probabilistic predictions based on a time-inhomogeneous queuing model, which can satisfactorily predict the availability of bicycles a few hours in advance and thus provide guidance to the end-user (Gast et al., 2015).

[2] Linear Regression, Gradient Boosting Machine, Support Vector Machine, Boosted Trees, and Extreme Gradient Boosting Trees

The travel planner software niche has been growing, so the interest in selecting the best travel routes, including those that are faster to do when using bicycles, is of great interest to scholars in this area (J. Wang et al., 2022).

Although the rebalancing of shared bicycle systems is one of the main factors included in the problem of the analyzed studies, the prediction of bicycle availability has also sparked interest in other branches (Yoshida et al., 2019). In this sense, the search for models and methods that better fit the data related to other problems inherent to BSS, such as frequencies and station-level analysis, is remarkable. Anyway, there is plenty of room for the literature to mature in the context in which BSS data are characterized by zero-inflation; that is, the count data has an excess of zeros.

To predict the trips that will occur between a station of origin and destination on a given date, Lucas & Andrade (2021) came across with data set in which only 4.3% of the target variable were different from zero, which means that is an excessive amount of zero counts, which has implications in the data modeling. Making efforts to deal with the significant limitations of earlier studies in tackling the excessive number of zeros in the dataset, Lucas & Andrade (2021) developed models based on a method that will fit zero-augmented data, specifically the hurdle model and the zero-inflated model. The application of the referred models surpassed the generalized linear models, being the hurdle model the best one based on the Akaike Information Criterion (AIC).

With only 6.3% of the data being non-zero, Kim & Cho (2021) also used a model built on the Zero-Inflated Model to study the origin-destination pairs of the BSS of Seoul in South Korea, once the data were zero-inflated. The Zero-Inflated Model developed here is a negative binomial constituted of a zero-inflated binary logit regression aiming to classify zero counts and a negative binomial model to manage the non-zero outcomes (Kim & Cho, 2021). However, it is important to note that not all studies (Ashqar et al., 2021) at the station level are faced with the problem of excess zeros, although they also need to use count models to make predictions.

When investigating the frequency of e-bike trips, Xu & Wang (2019) resorted to the Bayesian zero-inflated with a random effect modeling method to understand the variables contributing to e-bike trip frequency. The authors claim that the frequency count of the e-bike trips is inflated by zeros. Earlier studies suggested that zero-inflated models can be used to solve the presented situation of zero-inflated counting, once typical Poisson and negative binomial cannot fit the data (Xu & Wang, 2019). In addition, the results state that the likelihood of ratio test, a test to compare the goodness of fit of two nested regression models, shows that the

random effects significantly increase zero-inflated model fitness in indicating whether a person will use an e-bike (Xu & Wang, 2019).

Other authors (K. Wang et al., 2018) also used zero-inflated negative binominal models applied to different issues related to BSS. In this specific case, the goal was to estimate hourly trips for five age cohorts: younger millennials, mid millennials, older millennials, Generation Xers and Baby Boomers. The authors emphasize that the excess of zero in the frequency of trips is generated by processes inherent to human activity, such as, for example, not using bicycles at dawn or on days with fog and rain, and the zero counts that arise through to BSS infrastructure issues (K. Wang et al., 2018). The models developed include two components: the first for counting the frequency and the second one given by a probabilistic part dedicated to dealing with the zero use of bicycles, considering the impacts of time and weather. After analyzing the over-dispersion parameters, likelihood ratio test and Vuong Test, they concluded that zero-inflated negative binomial model could better estimate hourly trips for the five age cohorts (K. Wang et al., 2018).

Other authors (Soheil et al., 2020) also, facing the need to deal with a considerable number of zeros, chose to use generalized extreme value (GEV) count models instead of using zero-inflated or hurdle models. They argue that inflated frameworks models have a complex structural model that is difficult to estimate in general, especially when it is necessary to deal with other excesses, such as excess counts of ones, twos, and threes (Soheil et al., 2020). Through this study, it was possible to predict, with some success, the arrivals, and departures in an aggregated way, that is, for the entire bike-sharing system, but also in a disaggregated way, that is, for each of the bike racks (Soheil et al., 2020).

The literature on BSS is expanding, and considerable efforts have been made on one of the key issues of these systems: the need for an effective rebalancing of bicycles, providing users with the permanent possibility of starting a journey. However, as it was possible to verify throughout this literature review, in view of the organic flow of bicycle trips, it is difficult to guarantee that there are always bicycles available at all bicycle stations at a given time. For this problem, there have been repeated and diverse ways of forecasting demand and rebalancing the bicycles across the BSS stations. But while efforts continue to be undertaken to rebalance the bikes, users, when faced with empty bike racks, still have no information of any kind about when it will be possible to go on a ride.

A way to bridge this gap is by informing the user if bicycles are expected to arrive at a given station in the next few minutes. Nevertheless, the count of bicycles arriving at the stations has a zero-inflated distribution, which makes it difficult to predict how many bicycles will

arrive. Given these two gaps, it was intended to predict how many bicycles will arrive, for each of the stations and to examine whether it is possible to obtain more satisfactory results by addressing this question through ZIR.

CHAPTER 3

Methodology and Results

In order to understand whether it is possible, through zero-inflated data science techniques, to predict the arrival of bikes to a BSS station, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was applied. This methodology consists of a set of good practices, divided into six main steps: (i) Domain understanding; (ii) Data Understanding; (iii) Data Preparation; (iv) Modeling; (v) Evaluation; (vi) Deployment.

The CRISP-DM procedural model will help explore how data science techniques can inform the user if there is no bicycle available at a particular station and whether or not a bicycle will arrive in a specific time window, thus helping in improving the user experience.

3.1. Domain Understanding

Through the Systematic Review of the Literature, one learned that shared bicycle systems have been growing and consolidating themselves as complementary and essential transport modality in urban centers worldwide. However, for the proper functioning of these systems, several problems need to be dealt with efficiently, such as, the lack of organic balance in the distribution of bicycles in the stations to ensure availability. Nevertheless, in addition to the issue of redistribution of bicycles, an issue that has received due attention in the literature, other problems have not been widely addressed such as the lack of communication with the end user about whether a bicycle will be arriving at a given station or not. In this way, it is important to study whether it is possible, through zero-inflated data science techniques, to predict the arrival of a bike to a BSS station in the next few minutes.

This is a frequent question., precisely because of the lack of effective balancing, or even the overload of the system, stations often do not have any bicycles available, which prevents the user from starting a trip. However, the cyclist's experience could be improved if he had information about the bicycle’s probable arrival time at that station. Empowering the end user with this information the cyclist has satisfactory conditions to assess whether it is worth waiting for the arrival of a bicycle at the station, or if it is better to travel in another way.

Similarly, information about when it will be possible to start a new journey is quite common in different schedule-based types of transport, such as public buses, trains, boats, and

many others. Supplying information if it will be possible to start a bicycle journey - not only improves the user experience by ensuring predictability of the service status - but also brings Bicycle Share Services closer to more traditional means of transport in the sense of it being viewed as a mobility solution.

To bridge this gap in the literature, the goal of this work is to develop, from the cyclist's perspective, a proof of concept on the feasibility of informing the user about the possibility of starting a trip in a pre-defined time interval. Along these lines, it is intended to predict, for each station, whether a bicycle will arrive soon. If the forecast is positive, that is, if the model predicts that bicycles will arrive in in a while, it is also important to approximate how many bicycles will arrive. To achieve this aim, the technical goal of this work is to develop a predictive model to understand whether it is possible to predict the arrival of a bike to a BSS station in the next 60 minutes and to provide that information to the cyclist. If the indicator that a new bike will arrive is positive, it is still essential to predict how many bikes will arrive.

3.2. Data Understanding

The data analyzed in this study is the integration of three distinct datasets. The first one, which is the fundamental database, has the records of all trips that took place in the year 2021. A second database, also from Metro Bike Share System, is analyzed, but it holds descriptive information about each of the bicycle docks in the system, so it is essential to understand the properties of these docks, such as where they are located and how many there are. Additionally, meteorological data are analyzed and joined with the other two mentioned above, in order to hypothesize that exogenous variables can improve the understanding of the original data and provide better predictions.

3.2.1. Los Angeles Metro Bike Share System Exploratory Data Analysis

This work uses an open-source data from the Metro Bike Share System, a BSS in the metropolitan area of Los Angeles, California, USA.

The data with travel information were extracted from the Metro Bike System website[3], where datasets grouped quarterly with data from all trips that took place in 2021 are available. In this work, data from all the quarters of 2021 were extracted to be further analyzed, totaling 12 months of observations and fifteen variables.

The fifteen variables of this dataset are:

- **trip_id:** Locally unique integer that identifies the trip;

[3] https://bikeshare.metro.net/

- **duration:** Length of trip in minutes;
- **start_time:** The date/time when the trip began, presented in ISO 8601 format in local time;
- **end_time:** The date/time when the trip ended, presented in ISO 8601 format in local time;
- **start_station:** The station ID where the trip originated (for station name and more information on each station see the Station Table);
- **start_lat:** The latitude of the station where the trip originated;
- **start_lon:** The longitude of the station where the trip originated;
- **end_station:** The station ID where the trip terminated (for station name and more information on each station see the Station Table);
- **end_lat:** The latitude of the station where the trip terminated;
- **end_lon:** The longitude of the station where the trip terminated;
- **bike_id:** Locally unique integer that identifies the bike;
- **plan_duration:** The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up);
- **trip_route_category:** "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips;
- **passholder_type:** The name of the passholder's plan;
- **bike_type:** The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes;

The work was developed iteratively, and it was noticed that the computer processing power would have a high cost, therefore, would be necessary for high RAM abilities. This way, the Google Colab Pro+ license was used, in which 52GB of RAM was made available. These settings, however, were not enough for the study to be conducted with all the data available on the the Metro Bike Share System website; for this reason, only data from 2021 is used, which stands for 220.997 rows.

To get to know the data and its variables better, Exploratory Data Analysis (EDA) is carried out to discover patterns, detect anomalies, and verify assumptions. The analysis process was started by checking the completeness of the columns and if there were missing data, which would need to be handled. Missing data was detected in the following columns: *start lat, start_lon, end_lat, end_lon, start station name* and *end station name*. These last two columns were excluded from the travel log dataset as they were not common to all quarterly files. The treatment of other cases of missing data will be detailed in the following. When checking for duplicate trips based on trip ID, 17880 duplicate records were found. As duplicate data interferes with the EDA process, all duplicates were removed.

Next, the variables *day_of_week, start_month, start_year, start_date, end_date, start_hour, end_hour, start_minute, end_minute, business_day,* season, and *time_period* were

created from the existing variables. The creation of these variables is important because it eases the exploratory data analysis process, since it provides greater interpretability, and later, they may have a better predictive capacity about the variables from which they originated.

It was initially observed that, on average, during the year 2021, 556.46 trips took place per day. However, it is possible to observe in Figure 3.1, that there are sinuous movements concerning the number of trips that took place throughout the year. In October, for example, there was an abnormal peak in demand, so it is crucial to investigate whether it is possible to detect any explanation for this high demand and for the other patterns detected.

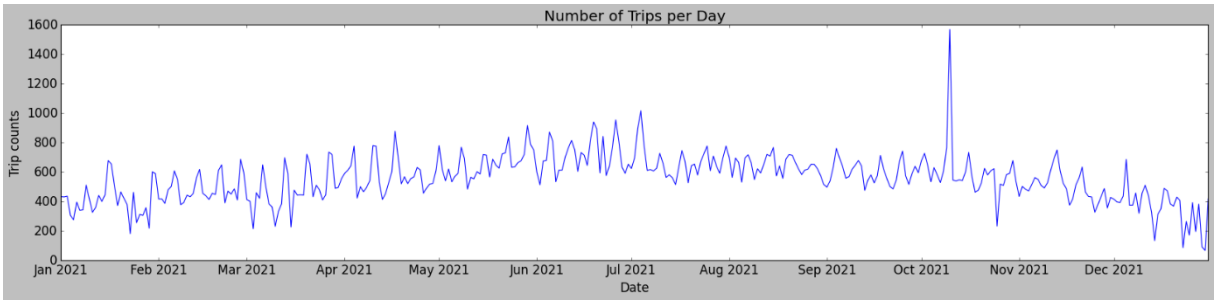

Figure 3.1 - Number of Trips per Day

When analyzing the boxplots of trips by month in Figure 3.2, it is possible to see an increase in the counts during the mid-year months. Attention is drawn to the high peak of trips that occur in October, but this is explained by the fact that *CicLAvia–Heart of LA* took place in that month, an event promoted by Metro in which the streets are closed to traffic and bikes ride is for free. On the other hand, the month in which there were fewer trips was December, particularly around the Christmas and New Year celebration dates.

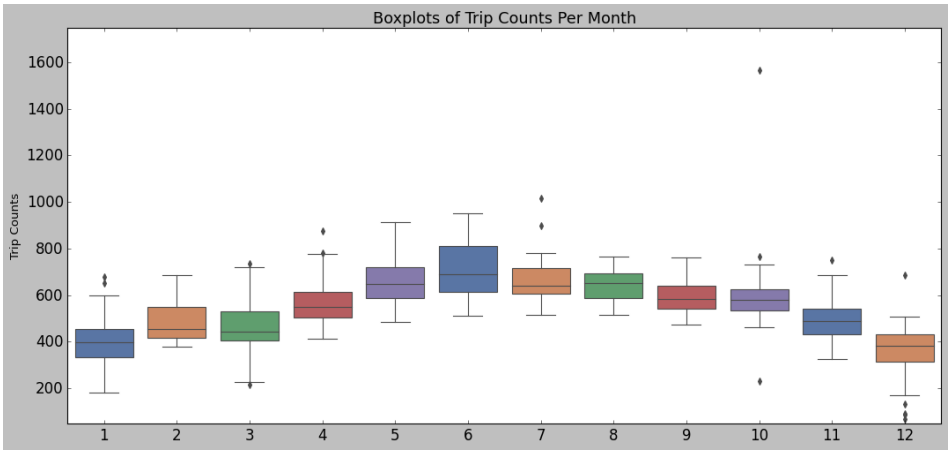

Figure 3.2 - Boxplots of trip Counts Per Month

To have a first idea about the characteristics of the data, descriptive statistics measures were calculated to understand each variable better. Through the analysis of these statistics, it was possible to notice, for example, that the duration variable has an average value far from the maximum value, so a closer look at this variable is important, as can be seen in Figure 3.3.

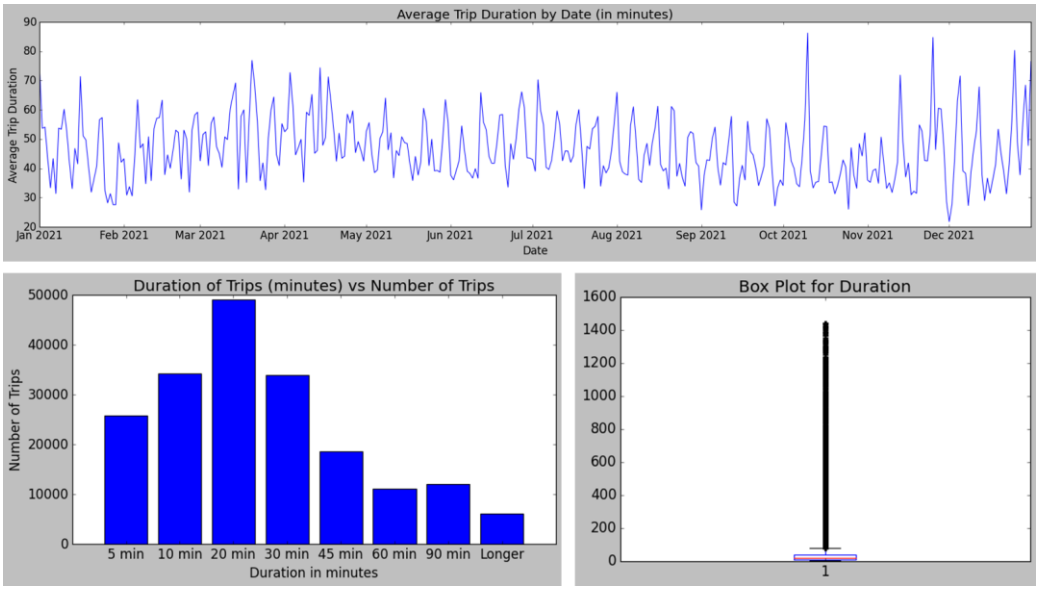

Figure 3.3 - Duration analysis

The high frequency of trips lasting more than 90 minutes is explained by the fact that, for all trips with technical problems, the system often does not compute the real end time of the trip so long trips may be the result of a system or user error.

Based on the geographic coordinates of the start and end stations of the trip, it was possible to calculate, through the Haversine Distance Formula - the distance between two points on a sphere - the angular distance between the start and end stations. In Figure 3.4 it is possible to see the distribution of trips by distance.

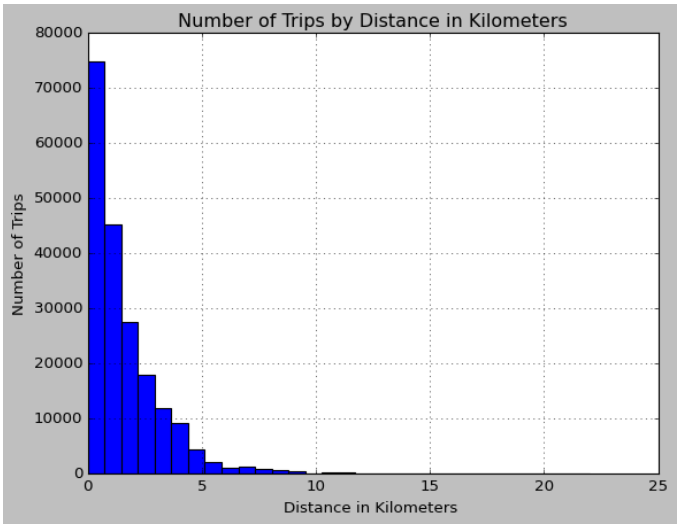

Figure 3.4 - Number of Trips by Distance in Kilometres

Note that as this feature is created based on the start and end stations and it is not intended to measure the distance cycled by an end-user, but the direct distance between the start and end of the trip, which in some cases may inform whether a cyclist has traveled to a near or a far station. Trips that start and end at the same point have a distance equal to 0, since it is impossible to associate any distance to that trip through the available data.

Through the Figure 3.5 it is possible to observe that the demand for bicycles has an increasing trend throughout the day. The demand starts to grow early in the morning and reaches its peak at 17:00, after which the demand drops dramatically. However, when analyzing the stratification for business days and non-business days, it is possible to conclude that demand is lower on non-business days and that the demand curve is flatter, which means that demand is more distributed throughout the day.

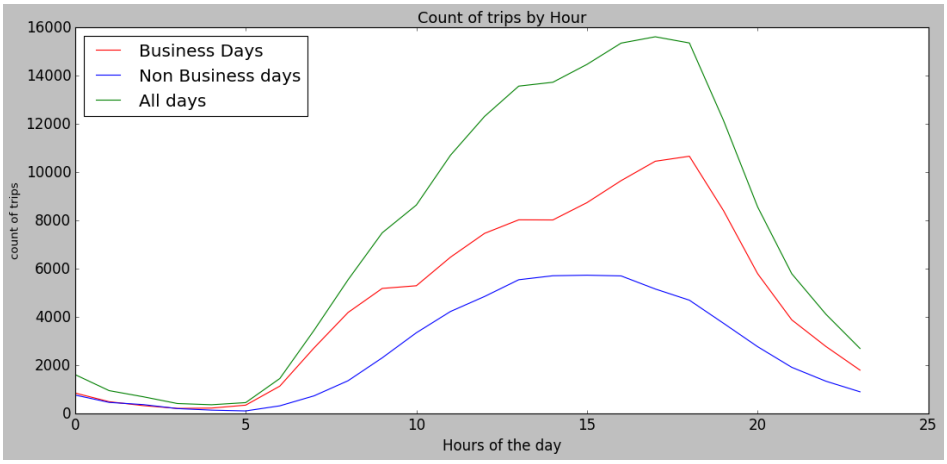

Figure 3.5 - Count of Trips by Hour

Before continuing to the subsequent analyses, it was decided to remove the remaining missing values, since these values reduce the statistical power and the next steps of the analysis require no missing values.

3.2.2. Stations Overview

In addition to the database with travel information, the database with details from all stations was also used and joined to the trip dataset to have more descriptive information on start and end locations.The four variables of this dataset are:

- **Station ID:** Unique integer that identifies the station (this is the same ID used in the Trips and Station Status data)
- **Station Name:** The public name of the station. Staff uses “Virtual Stations” to check in or out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station.
- **Go live date:** The date that the station was first available
- **Region:** The municipality or area where a station is located
- **Status:** "Active" for stations available or "Inactive" for stations that are not available as of the latest update

In the dataset with travel records for the year 2021, 217 stations were spread over three distinct regions: Downtown LA, Westside, and North Hollywood, as seen in Figure 3.6.

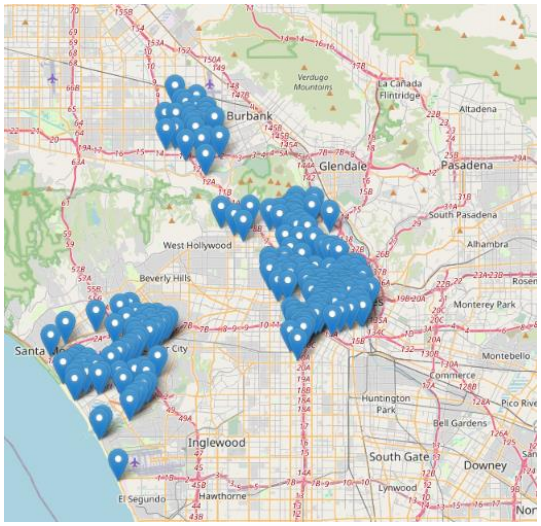

Figure 3.6 - Location of stations in the city of Los Angeles

It is worth noting, through figure 3.7, that even though not substantially different, the bike stations with the most demand for a start-trip vary between non-business days and business days.

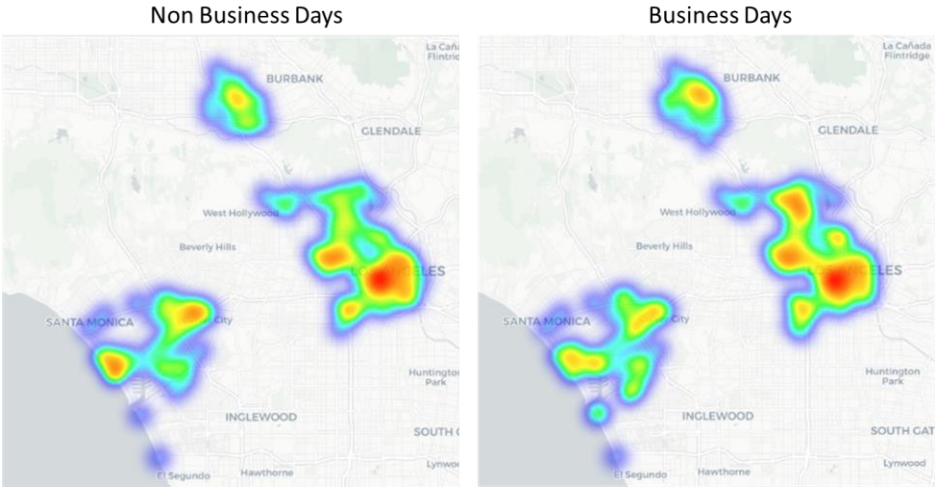

Figure 3.7 - Heatmap comparing bike demand on business days and non-business days

Regardless of whether it is on a business day or not, the stations in the downtown area are the most demanded. However, it is possible to note that there is a variation in the stations that are the most popular in Downtown Los Angeles depending on whether it is a business day or not. As far as West Los Angeles is concerned, demand in this region increases during non-business days, which may be explained by the proximity to beaches and other leisure areas. However, to confirm this, it would be necessary to develop a study of the points of interest in the city of Los Angeles and how the activities of these points impact travel demand

During the analysis of the stations, it was observed that the popularity of the stations has a standard deviation of approximately 982.54, which means that the count of trips that starts in each station is not homogeneous, thus being dispersed from the average value 808.7. That said, it was found that the least popular bicycle station to start a trip during the year 2021 - Universal City station presented only four trip starts. The most popular station - Ocean Front Walk & Navy - presented 7836 trip starts. The frequency of departure in these two stations has a difference of 199.796%.

3.2.3. Adding Context Awareness: Weather Data

To add context awareness to this study, data from other sources were further combined in order to expand the scope and interpretative character of the analysis. Several other context information could have been integrated here, such as: real-time traffic information, data from episodic social gathering events happening in the city, and other situations that affect, in one way or another, how people travel. Nevertheless, for lack of resources, this work focuses solely on meteorological data combined with the travel dataset presented before.

The meteorological data were purchased through the Open Weather Map website[4], which offers a product called History Bulk, allowing the extraction of hourly historical weather data for over 40 years for any chosen location or pair of coordinates. It includes fifteen weather parameters represented in twenty-six variables. In this sense, History weather bulk for Los Angeles from January 01, 2016, until March 21, 2022, was obtained.

The twenty-six variables obtained are:

- **city_name:** City name
- **lat:** Geographical coordinates of the location (latitude)
- **lon:** Geographical coordinates of the location (longitude)
- **main.temp:** Temperature
- **main.temp_min:** Minimum temperature at the moment. This is deviation from temperature that is possible for large cities and megalopolises geographically expanded (use these parameters optionally).
- **main.temp_max**: Maximum temperature at the moment. This is deviation from temperature that is possible for large cities and megalopolises geographically expanded (use these parameters optionally).
- **main.feels_like:** This temperature parameter accounts for the human perception of weather
- **main.pressure**: Atmospheric pressure (on the sea level), hPa
- **main.humidity:** Humidity, %
- **main.dew_point:** Atmospheric temperature (varying according to pressure and humidity) below which water droplets begin to condense and dew can form. Units – default: kelvin
- **wind.speed:** Wind speed. Units – default: meter/sec
- **wind.deg:** Wind direction, degrees (meteorological)
- **wind.gust:** Wind gust. Units – default: meter/sec
- **clouds.all:** Cloudiness, %
- **rain.1h:** Rain volume for the last hour, mm
- **rain.3h:** Rain volume for the last 3 hours, mm
- **snow.1h:** Snow volume for the last hour, mm (in liquid state)
- **snow.3h:** Snow volume for the last 3 hours, mm (in liquid state)
- **weather.id:** Weather condition id
- **weather.main:** Group of weather parameters (Rain, Snow, Extreme etc.)
- **weather.description**: Weather condition within the group
- **weather.icon:** Weather icon id
- **visibility:** Average visibility, metres. The maximum value of visibility is 10km
- **dt:** Time of data calculation, unix, UTC
- **dt_iso**: Date and time in UTC format
- **timezone:** Shift in seconds from UTC

[4] https://openweathermap.org/

Another crucial step in this study is the process of adding context awareness to the trip data, making it possible to understand how external factors affect the observed trips. For this work, the focus will solely be meteorological data, as these are easy to obtain and seems to have a profound impact on travel frequency, hence, they are candidates to be good predictors for the count variable (B. Chen et al., 2013; Lucas & Andrade, 2021). It is emphasized, however, that the combination of data from multiple sources can be a factor capable of enriching the predictive models; for this reason, several other types of external data could be combined here to assess this increase in the quality of forecasts, as highlighted in the literature review (Cerqueira et al., 2021).

After reading the CSV file, procedures to transform the date format were applied so that it was possible to integrate the hourly weather data with the corresponding date and time in the trip records data frame. Then, variables whose values do not vary or data were unavailable were excluded, since the lack of assigned values and variables with only one frequency, do not bring relevant information in the forecasting aspect. The list of the excluded variables is the following: *snow_1h, snow_3h, grnd_level, sea_level, city_name, lat, lon, weather_icon, dt, dt_iso.*

All sets of data were merged into one. After merging, it was necessary to assign values equal to zero for situations where NA (not a number) was present in the variables *rain_1h, rain_3h* and *wind_gust*, because in this specific case, the NA represented that no rain or gust had been observed.

As one can see in Figure 3.8, most trips took place when the sky was clear. However, there were also many trips when it was cloudy and with dry fog.

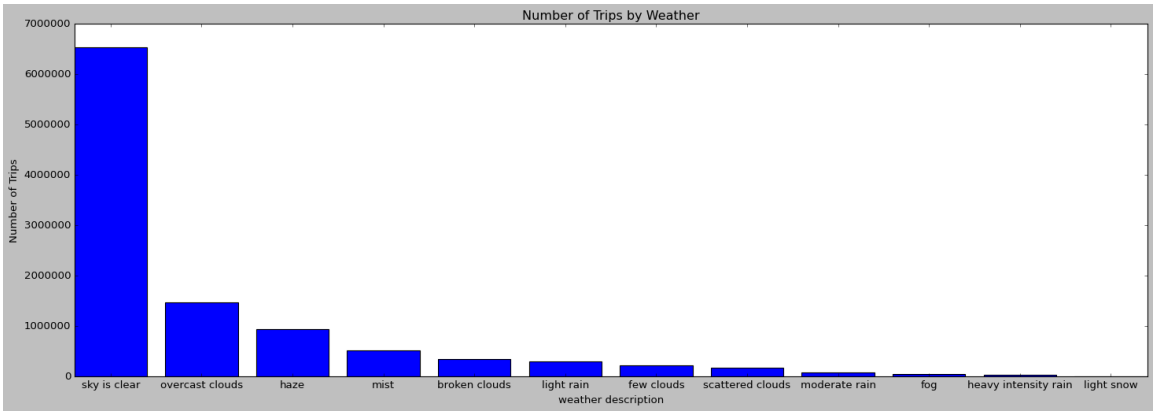

Figure 3.8 - Number of Trips by Weather

Regarding the temperature at which the trips took place, most trips occurred when the temperature was between 15 and 20 degrees. However, it is possible to observe, through figure 3.9, that some trips also took place when the temperature was extreme, that is, close to 5 degrees or close to 35 degrees.

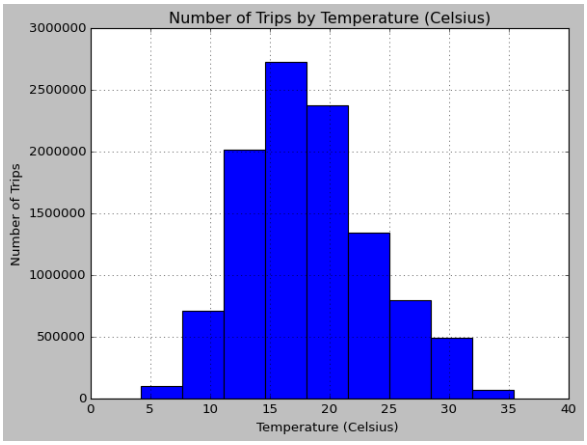

Figure 3.9 - Number of Trips by Temperature (Celsius)

3.3. Data Preparation

In this step, the goal was to fix anomalies and prepare the data for developing a prediction model. Thus, the aim is to transform and adjust the input data so that it is possible to have a high-quality and adequate dataset to use in training the model.

Metro Bike Share highlights that a *virtual station* is used "by staff to check in or check out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station" (Metro Bike Share, 2022). Thus, all trips that started or ended at the *virtual station* were excluded from the records once it was intended to analyze the trips that took place in an organic way.

As expected, due to the previous descriptive analysis of the variables observed during the data understanding, through the Interquartile Range Rule (IQR) relating the box plot presented in Figure 3.3, it was possible to confirm that the *duration* variable presents a huge number of outliers that it was decided to remove, as shown in figure 3.10.

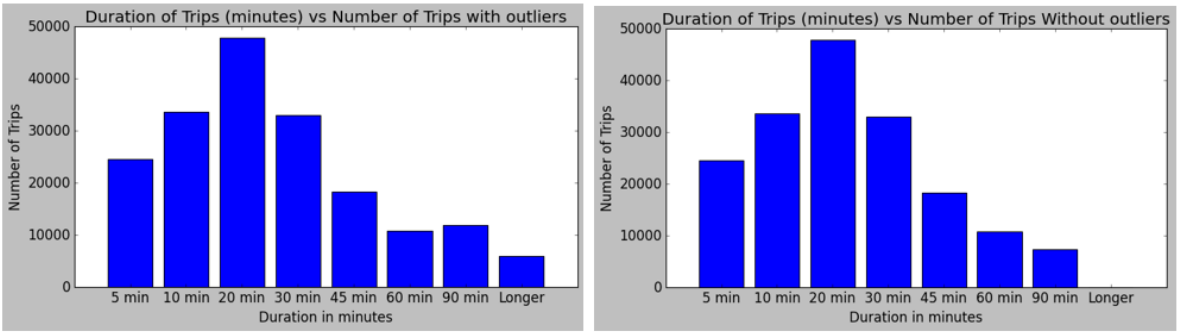

Figure 3.10 - Comparison of trip duration data with and without outliers.

3.3.1. Establishing relevant time intervals

This step in the Data Preparation is dedicated to creating the data structure that will be used in the forecasting models. The focus of this work is to give some predictability to the cyclists about how the situation of not having any bicycles in each station would evolve. Initially, it was thought that supplying a forecast of how many bicycles would arrive at a given station in a 10-minute interval would be a practical choice, as, in comparison with other more traditional means of transport, such as bus, subway, and train, 10-minutes intervals are in general an acceptable waiting time till the transport arrives. However, there is a trade-off when it comes to setting up these time intervals. The smaller the time interval, the greater the volume of data, in view of the resampling to be performed. In other words, using 10-minutes intervals requires a higher computational capacity than a 60-interval window. In addition to these difficulties, by using a shorter time interval, the complexity of the modeling process increases substantially. Therefore, as the objective is to provide the user with information about whether it will be possible to start a trip, in the next few minutes and verify if the Zero-Inflated Model can predict whether new bicycles will arrive at a given station, the 60-minute interval will be used, which means that the constructed analysis works as a proof of concept in order to verify how many bicycles will arrive the next hour.

To make this forecast available based on 60-minute intervals, resampling will be necessary since the data are not available at the same frequency in which predictions are intended. As seen in Figure 3.11, each row of the original dataset represents an individual trip. The following procedures were followed: (i) change in the frequency of the observations into 60-minutes intervals; (ii) create a new column called "*count"* to count how how many trips took place in each of the 60-minute interval.

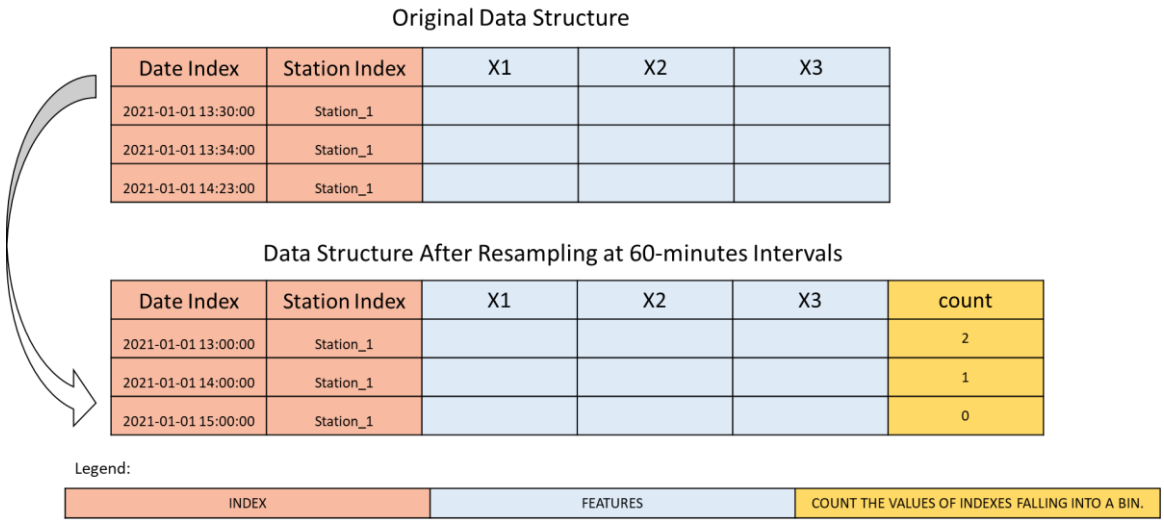

Figure 3.11- Change in data structure after resampling process (author's elaboration)

With this grouping, individual characteristics of the trips had to be disregarded, such as the indication whether the bicycle used in the trip is electric or mechanical, what type of contract the cyclist has with the Metro Bike Share System, if the trip is round-trip or one-way and all the characteristics that are related to the beginning of the trip,. Since the objective is to analyze the frequency with which bicycles arrive, regardless of which station they departed from, the departure station and the attributes that detail the beginning of the trip or the client are irrelevant.

On the other hand, the grouping was created according to the station where the bicycle was delivered and when it was delivered. Since the stations started working on different dates, the resampling process must be done individually for each one of the stations.

Another particularly crucial point is that, during the resampling, there are intervals in which no trips arrived, meaning the count for these intervals will be equal to zero. Since these intervals are more frequent that the intervals in which a bicycle arrived, the resampling process inflates the count variable with zeros, gaining a new characteristic: is zero-inflated, which means that the frequency of zeros is excessively high compared to the frequency of other values.

During the resampling process, variables such as *Distance_KM* and *duration*, as they are features that describe the trip, were also aggregated. Thus, selecting a summary statistic to calculate the new aggregated values is necessary. The average *Distance_KM* and *duration* for the trips that occurred in the respective interval was assigned to each of the grouped rows.

It was also necessary to assign values equal to zero for the distance and duration in the cases of intervals where no trip was observed, since, when resampling, these variables had unassigned values, because there were no observations to perform the average.

As the objective is to predict how many bicycles will arrive in the next 60 minutes, it was necessary to create a new target column, called *next_count*, which, as the name implies, is the result of shifting the newly created *count* variable so that the *y* of a given row is the count of bicycles that arrived in the subsequent 60 minutes, defining this way the target variable of this study.

3.3.2. Feature Selection

This step is dedicated to the process of selecting variables that are relevant to the development of the model. The selection of variables aims to, through statistical tests, to select only those variables with the ability to predict the target variable.

This process began by transforming string variables into categorical ones so that it was possible to study the correlation between them, as seen in Figure 3.12.

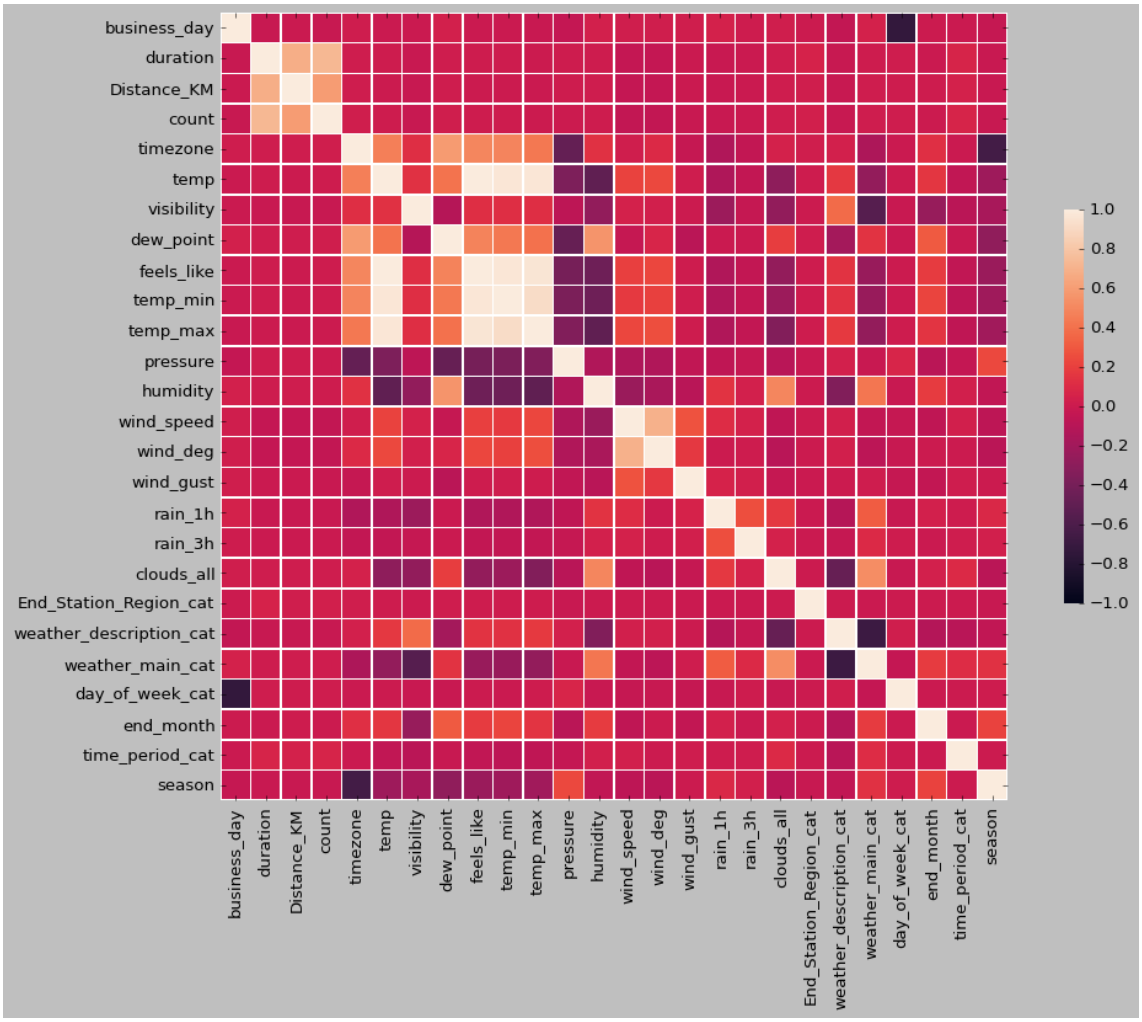

Figure 3.12 -Correlation Heatmap

Firstly, it was found that *visibility* has a high degree of correlation with the *weather_main_cat* variable. However, the *visibility* variable presents null values preventing its use in the following processes. Furthermore, as the correlation between these two variables shows a value between moderate and high, it was decided to exclude the *visibility* variable to the detriment of dropping rows with NA values. As several other variables present high correlation values it was defined that, for each pair of features correlated by a value greater than 80%, the variable with the lowest correlation with the target variable should be eliminated. By applying this logic, three variables were eliminated: *temp_min*, *temp_max* and *temp*.

In addition to the study of correlation, which is always applied to a pair of variables, it is also interesting to conduct the study of multicollinearity, which is the verification whether there is a linear relationship between three or more independent variables, even if no pair of variables has a high correlation. Detecting multicollinearity is important because it reduces the statistical significance of the independent variables. The Variance Inflation Factor (VIF), a measure of the amount of multicollinearity, was calculated However, when calculating the VIF for each of the features, a threshold must be set to select the variables according to the value of the VIF. And the question that arises here is the following: what is the best threshold value of VIF? Some authors argue that a VIF of 2.5 or more can already indicate considerable collinearity (Johnston et al., 2018); other scholars are more tolerant and would consider using a VIF lower than 10 (Allison1999, p. 142 *apud* Johnston et al., 2018). As a measure of balance, variables presenting VIF lower than 5 in the Table 2 were selected for the modeling step.

Table 2 - Features and their respective variance inflation factor (VIF)

VIF VALUE	FEATURE NAME
1.077	rain_3h
1.108	wind_gust
1.284	rain_1h
1.473	End_Station_Region_cat
2.045	Distance_KM
2.332	count
2.513	clouds_all
2.545	wind_deg
2.840	duration
3.144	wind_speed
3.152	time_period_cat
3.580	weather_main_cat

5.005	end_month
7.017	day_of_week_cat
7.095	business_day
12.337	season
17.744	weather_description_cat
83.529	dew_point
184.818	feels_like
239.409	humidity
744.060	timezone
1226.982	pressure

3.3.3. Preparation for Zero-Inflated Framework

This study aimed to predict how many bikes would arrive at a given station using zero-inflated techniques. Zero-Inflated Models are considered metamodels as they depend on other estimators to make predictions. The Zero-Inflated Model is a meta-regressor for zero-inflated datasets, that is, when the target variables have a lot of zeroes and two tasks must be performed: classification and regression. The classification task aims to classify a binary variable, in this case, are bikes arriving in the next 60 minutes or not. The regression tasks deal with all non-zero values predictions and are to be used only after the classifier determines that the predicted value is not a zero. In summary, a Zero-Inflated Model first classifies whether the output is zero. If yes, then it outputs zero. Otherwise, the regressor goes into action, performs its prediction, and then outputs it.

Hence, it is necessary to arrange the data of this work in a data structure based on a time series, so that features that characterize past trips contribute to the prediction of future bicycle counts arriving at a station. It is important to note that the Zero-Inflated Model regression task, in the context of the *scikit-lego* library, technically requires it to be a regressor of the *scikit-learn* library. In short, by arranging the dataframe in a time series structure and treating the goal of this work as a supervised learning problem, it will be possible for the time series prediction to be performed within the scope of the Zero-Inflated Model regression task.

3.4. Modeling

As previously referred, the goal of the modeling step is to develop and fit a model capable of predicting how many bicycles will arrive in the next 60 minutes at a given station using the data prepared in the previous steps. It is important to note that since it is intended to make forecasts individually for each bike rack, the modeling process is applied separately for each one of the stations, considering that, since behaviors are different between the stations, there is a different

time series for each one. The modeling process consists of the following four steps (Figure 3.13), which are triggered apart for each of the bike stations in a *for-loop statement*:

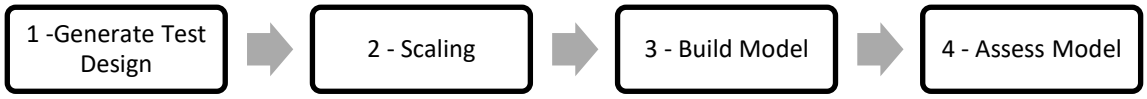

Figure 3.13 - Modelling Process Steps

During the first step of the modeling process consists in the generation of the test design: the division of the data set into a training and test set, where the former will be used to train the model, and the latter used to test whether the training was successful. Since the data in this study are time series it is mandatory to have samples that are observed at fixed time intervals, *TimeSeriesSplit* function, a time series cross-validator from *Sklearn*, is implemented to split the data set into training set and testing set while maintaining data continuity and, in parallel, carry out a walk-forward cross-validation behavior. Still, it is important to define the sample size used in the training set. If not indicated otherwise, the training set will automatically consist of the number of samples divided by the number of splits + 1. Therefore, all models will be trained with five different train-and-test designs to find the optimal split size: (i) 90%-10%, (ii) 85%-15%, (iii) 80%-20%, (iv) 75%-25% and (v) 70%-30%.

During the second step, called scaling, to make it easier for a model to learn and understand the problem, the values of each one of the features were scaled. The third step, where the model is built, is intended to define the model algorithm and its parameters and describe the model to guarantee that the results can be interpreted. The last step is the one that evaluates the model. Depending on the evaluation results, the parameters are reconfigured (in step three) to improve the model's performance.

By implementing these steps individually for each of the stations, each one will have its own prediction model, in which the learning process will consider its particularities. Furthermore, as information on how the situation on bicycle arrivals will be given at the station level, it is important to understand the intrinsic patterns of bicycle racks in order to create models capable of generalizing the specific activity of each one. The aim is to create models and progressively increase their complexity. In this sense, the following models were built within the Zero-Inflated Model framework, which, as mentioned before, performs a classification and a regression task.

3.4.1. Zero-Inflated Regression Models

The ZIR Models were built considering only the variables that have a VIF lower than 5, calculated after eliminating highly correlated variables; each model was cross validated five times. In order to reduce the complexity and to develop a proof-of-concept of this methodology, the models presented in the Table 3 were built considering only the North Hollywood region of the City of Los Angeles. It should also be noted that the models were developed considering the default configuration of their hyperparameters.

Table 3 - ZIR models

MODEL #	CLASSIFIER	REGRESSOR
ZIR 1	Decision Tree	Decision Tree
ZIR 1.2	Decision Tree (Ada Boost)	Decision Tree
ZIR 2	Random Forest	Random Forest
ZIR 3	Extra Trees	Extra Trees
ZIR 4	Gradient Boosting	Gradient Boosting
ZIR 5	Gaussian Naive Bayes	Gradient Boosting

The first model developed in the Zero-Inflated Model is the ZIR 1 model, which is based on the Decision Tree algorithm. The models developed based on decision tree algorithms set up decision nodes that are related to each other by a hierarchy composed of the first node of the tree, which is the root node (represented by the features of the dataset) and leaf nodes, which are the results that the model generates. Additionally, a variation of ZIR 1 was built, the model ZIR1.2, an Ada Boost algorithm based on Decision Trees, which is an algorithm that builds a model using training data and builds the second model to correct the error that occurs in the first model, boosting the classifier (Choudhary & Narayan Singh, 2020).

After constructing the ZIR 1 and the variant with the AdaBoost model, it was decided to develop the ZIR 2 model based on Random Forest. This approach is designed following the logic that a single decision tree, done in the ZIR 1 model, may not be enough to produce satisfactory results, so the random forest algorithm will be implemented, since it combines the result of several trees created randomly, to output the result. In other words, random forests are a collection of decision trees, which are integrated into a single final result. Another expected benefit of ZIR 2 compared to ZIR 1 is that as random forests, compared to decision trees, use multiple trees and each tree is differently composed, so the chance of overfitting decreases, which is a significant improvement, considering that one of the main disadvantages of the decision tree is precisely its prone to overfit (Jun, 2021).

An important attribute of the models built based on the random forest algorithm is that these models, in their default settings, use a sampling process called bootstrap to build each of the trees. The Bootstrapping process reduces the variance of the prediction as it randomly chooses which data each tree will use from the training dataset. That is, in a dataset composed of distinctive features, the ZIR 2 model will sample subsets of observations with different features of the dataset. A decision tree is built on this subset (Jun, 2021).

However, it is important to observe how a forest of trees would behave using the whole original sample to construct each tree, so the ZIR 3 model, based on Extra Trees, is developed. Like Random Forest models, Extra Trees models also build several trees (a forest) to make predictions (Geurts et al., 2006). However, models based on Extra Trees, in addition to not doing bootstrapping, divide the node randomly, and the random forests calculate what would be the best division to be performed before dividing a node. Given these two characteristics of models based on Extra Trees, one expects that the ZIR 3 model will have a lower bias, since the entire data set will be used to construct a tree, however, it is also expected that the ZIR 3 model will reduce variance, as the node split point will be selected randomly (Geurts et al., 2006).

Both Random Trees and Extra trees models build their forests through independent trees, so the strengths and weaknesses of each of these trees are not analyzed. In this sense, to try to minimize the deficiencies of each of these trees, the ZIR 4 model was developed based on Gradient Boosting algorithm. In other words, unlike the forests developed in the ZIR 4 and ZIR 3 models, the ZIR 4 collaboratively builds the forest, that is, trees are gradually created to populate the forest, however each tree is designed to improve the existing weaknesses in the forest, in order to drive improvements concerning to the deficiencies of the trees that were produced previously. That process reduces the gradient of the loss function and therefore, it is expected that this method will yield better results than the previous models (Jun, 2021). although it is more prone to overfitting than models relying on Random Forest, for instance.

In light of the fact that the Naive Bayes algorithm is one of the most efficient and effective for solving machine learning classification problems (H. Zhang, 2004), ZIR 5 was developed because it is an easy-to-implement algorithm with competitive results compared to more complex machine learning models. Additionally, this algorithm is known to have a speedy training process (H. Zhang, 2004). However, although it is a reputable classifier, Naive Bayes is not a good estimator (H. Zhang, 2004), so the regression task for the ZIR 5 Model is based on Gradient Boosting. That said, the ZIR 5 is a hybrid model that combines a classifier with a

good reputation and a regressor with several strengths, as highlighted above, so satisfactory results are expected from this combination expressed through the ZIR 5 model.

3.4.2. Zero-Inflated Regression Models built Without Success

For a Zero-Inflated Model to be successfully constructed, the classifier must be able to predict that some training labels are non-zero. This prerequisite is essential for ZIR models to perform their two tasks: classification and regression, as regression is only performed if the classification predicts non-zero labels. In other words, if the classification task predicts that the training labels are all zero, the regression task of a ZIR model will become obsolete, so the classifier must be improved or changed for the regression to be workable.

It is worth noting that, during the modeling stage of this work, several classifiers were used to build ZIR models, but many were unable to predict non-zero labels and, therefore, could not be considered for comparison between ZIR models and plain regression models. This comparison is important for this work, as it is a way to assess whether the Zero-Inflated Model framework can more accurately predict whether a bicycle will arrive in the next 60 minutes.

ZIR models with the classification task based on Support Vector Machines (SVMs) had a tough time doing non-zero classifications. When using the standard parameters of the SVM classifiers, these models were not able to classify any non-zero label, so this algorithm, with its default settings, was unsuccessful for this work. Nonetheless, when defining the regularization parameter C with a remarkably high value, that is, above 230, and the gamma parameter, a parameter that controls the influence of features on the decision boundary, above 5, the model was able to classify labels different from zero. Assigning high values for the C parameter leads to overfitting, resulting in low bias and high variance. Therefore, more careful analysis and a thorough tuning process must be carried out, such that there is no overfitting (Ethridge et al., 2010). Thus, the ZIR models that guide the classification task in the SVM algorithm will not be considered for evaluation and comparison with the other ZIR models.

Like ZIR SVMs, Logistic Regression models within the ZIR framework was also unable to perform non-zero classifications with its default settings, which means that the model is obsolete, as it does not perform the regression task. Even after a hyper parameterization process, the ZIR model based on logistic regression continued to classify all values as zero, so this classifier also lacks a more careful tuning process that would make it comparable with other models or to draw conclusions that this model is ineffective for the problem that this work studies.

As far as some other classification algorithms are concerned, such as k-nearest neighbors vote, Gaussian Process Classification (GPC) based on Laplace approximation, Multi-layer Perceptron and Quadratic Discriminant Analysis, there has been an attempt to implement them as classifiers in the Zero-Inflated framework, but further analysis and development around the sample weight argument is required to successfully fit the model. In other words, Sample weight is an unexpected argument for the Zero-Inflated framework, so it was not possible to fit the previously mentioned models.

3.5. Evaluation

The objective of this section is to compare the performance of different models, more precisely whether models were able to predict how many bicycles will arrive in the next 60 minutes for each of the stations and to assess if models within the Zero-Inflated framework perform better than a plain regression.

3.5.1. Evaluation Metrics

Root mean square error (RMSE) - as an evaluation metric, is often better at identifying differences in the model performance since it gives higher weight to poor conditions(Chai & Draxler, 2014). The lower the RMSE, the better a model fits a dataset (Amalia et al., 2021).

Understanding the RMSE metric is more straightforward since the error is presented in the same unit as the response variable, because squared errors have units of their squared response, and the square root of the squared error is also represented in the same unit as its response (Kambezidis, 2012). In this study, since the response variable is the bike count arriving at a station in the next 60 minutes, the RMSE stands for the error between the observed and predicted bike count. In other words, if the RMSE is 0.5, for example, the forecast model is miscounted by half a bicycle.

Of note, it is important to reinforce that, regardless of the choice of metric for evaluating model performance, a single metric emphasizes only a certain aspect of the error characteristics, so a combination of metrics, such as RMSE and mean absolute error (MAE), for example, is a clever way to get a complete picture of how comparatively the models are performing. Theoretically, the RMSE score will never be smaller than the MAE score. This is due to how both metrics are calculated (Chai & Draxler, 2014). Moreover, a joint version of the two scores would not reproduce an analytically meaningful result (Hodson, 2022).

MAE, as the name suggests, is the mean absolute error that the model's predictions have in comparison with their corresponding actual observed values (Hodson, 2022). This means that

the closer the MAE value is to zero, the better the model predictions are. Thus, the RMSE and MAE metrics simultaneously will be used to evaluate whether the models built based on the Zero-Inflated framework are more proper to answer how many bicycles will arrive in the next 60 minutes, if any.

3.5.2. How are RMSE and MAE computed?

As described earlier, there were five different training-test splits for all the models developed in this work. Each training and test set created for every station was cross validated five times. The score of each training-test split is the average value of the five cross validation interactions.

At this point, every station has five RMSE and five MAE computed, but it is necessary to choose one score for each metric to make the comparison across the models easier. Thus, the training-test split that resulted in the RMSE and MAE closest to zero, but greater than zero, is chosen as the final score for a given station.

After the mentioned procedures, each station has only one score for the RMSE and only one score for the MAE, so that the calculation of these metrics for the overall model gets simpler. The global metrics of each of the forecast models are given by the average of the scores of each of the stations. That is, each of the models has only one RMSE and one MAE.

3.5.3. Baseline

The evaluation process of the Zero-Inflated models developed in this work begins with defining a baseline model. This is a reference model that will inform - before a detailed analysis of the results - whether the objective of this work is being attained. The baseline model can thus be depicted as a plain decision tree regressor, a decision tree model that was not built within the Framework of the Zero-Inflated Model.

This baseline considered all the data preparation presented earlier and followed the entire model building method as reported in the CRISP-DM modeling step.

3.5.4. Comparing the developed models

This section aims to assess whether, by framing the problem in the Framework of the Zero-Inflated Model, it is possible to reduce the RMSE and MAE of the developed models. To compare the performance of the Zero-Inflated Model framework - for the models developed through this method - analogous models were also developed: Plain RF, which means Plain Random Forest Regression and Plain GB, which means Plain Gradient Boosting Regression.

Analog models are single regression models using the same regression algorithm that a ZIR model was based on, but without being within the Zero-Inflated framework.

As shown in Figure 3.14, all ZIR models have a better performance, in terms of RMSE, than the models not included in this framework, although they used the same regression algorithm.

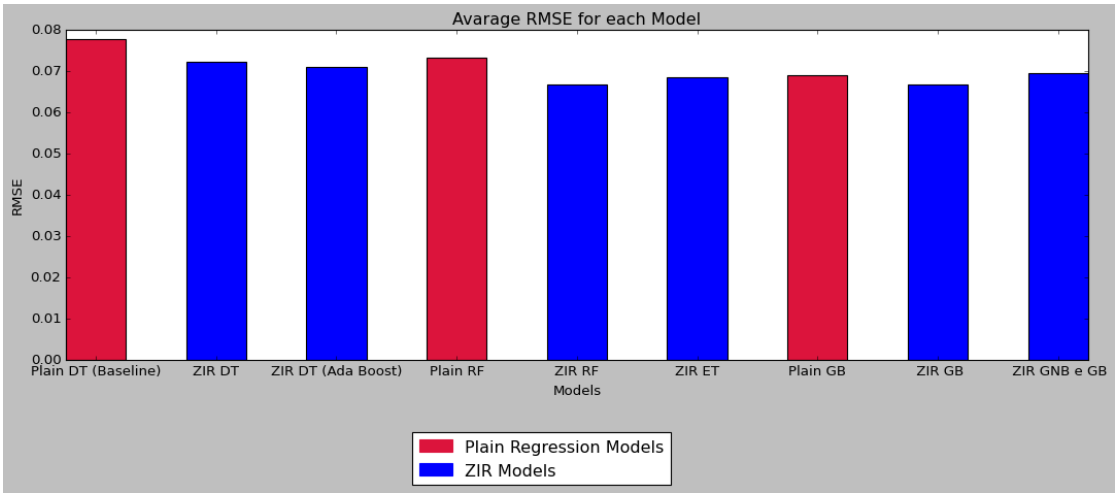

Figure 3.14 - Average RMSE for each Model

The improvement observed in the ZIR models over the analogous regression models can be explained by the classification task that the ZIR models conduct before carrying out the regression assignment. By performing the classification task before starting the regression, the classifier reduces the complexity of the task that the regressor must do, since the regression will not need to deal with such an excessive number of zeroes.

In Table 4 it is possible to observe how much the ZIR models were able to minimize the RMSE and MAE.

Table 4 - Models developed and their respective scores

MODEL #	CLASSIFIER	REGRESSOR	$\bar{x}$ RMSE	$\bar{x}$ MAE
BASELINE	Plain Decision Tree Regressor		0.0777	0.0173
ZIR 1	Decision Tree	Decision Tree	0.0721	0.0068
ZIR 1.2	Decision Tree (Ada Boost)	Decision Tree	0.0701	0.0066
PLAIN RF	Plain Random Forest Regressor		0.0733	0.0177
ZIR 2	Random Forest	Random Forest	0.0669	0.0060
ZIR 3	Extra Trees	Extra Trees	0.0686	0.0063
PLAIN GB	Plain Gradient Boosting Regressor		0.0690	0.0182
ZIR 4	Gradient Boosting	Gradient Boosting	0.0667	0.0061

ZIR 5	Gaussian Naive Bayes	Gradient Boosting	0.0693	0.0179

In relation to the Baseline, the ZIR 1 model reduced the RMSE by 7.21%. When the Ada Boost Algorithm (ZIR 1.2) is applied to the Decision Tree Classifier task, the reduction of the RMSE over the baseline is approximately 9.78%. Regarding the models developed using the Random Forest (RF) algorithm, ZIR 2 outperformed Plain RF by reducing the RMSE by 8.73%. If the comparison is between ZIR 2 and the Baseline, the RMSE reduction reaches 13.90%. As for the models based on Gradient Boosting, it is seen that the Model ZIR 4 also outperformed its Plain version, although the reduction of the RMSE, in terms of percentage, was more modest than the algorithms compared previously. Notwithstanding it is not such a significant difference in terms of percentage, it was still possible to reduce the RMSE of Gradient Boosting by 3.33% while framing it within the Zero-Inflated framework. Nonetheless, ZIR 4 has an RMSE 14.15% lower than Baseline.

An explanation for why the difference between the RMSE for the Plain GB model and the ZIR 4 model is not highly significant is that the Gradient Boosting algorithm is prone to overfit, and the Plain GB Model used the parameters by default and therefore the number of trees added to the model was not limited enough to prevent overfitting. However, such a hypothesis would need to be confirmed through tuning, that is, selecting proper hyperparameters. The ZIR 5 model, on the other hand, performs 10.55% better than the baseline, however, when comparing its performance to the Plain GB model, the ZIR 5 model, with its standard hyperparameters, failed to reduce the RMSE.

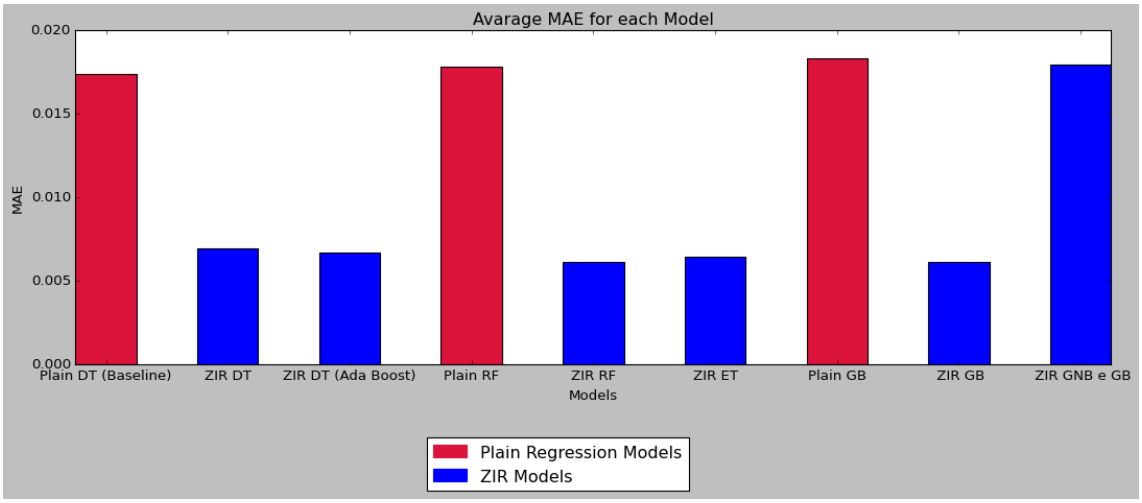

Figure 3.15 - Average MAE for each Model

When comparing the models using MAE score, as seen in the Figure 3.15, the conclusion is that the ZIR models outperform the plain regression models. For example, by comparing ZIR 1 model to the baseline, ZIR 1 has a 60.69% reduction in MAE score. The reduction in MAE becomes even more remarkable when contrasting the performance of the ZIR Grandient Boosting based model. There was a 66.48% reduction in the MAE of ZIR 4 over Plain GB. Nevertheless, it is important to pay careful attention to the characteristics of the MAE and RMSE before drawing any conclusions about the benefits of ZIR models over simple regression models.

As mentioned earlier, RMSE penalizes outlier errors more strongly than MAE. That is, when analyzing the performance of the models developed using MAE, it can be seen that the simple regressions err at a much higher frequency than the ZIR models. On the other hand, when analyzing the error of the models from the perspective of RMSE, it is observed that although the plain regressions comment errors more frequently (conclusion drawn with MAE), the magnitude of the error is similar in all models.

Once RMSE and MAE are averaged forms of the L2 norm and L1 norm, which are the Euclidean and Manhattan distance, respectively (Hodson, 2022), RMSE ignores small errors and amplifies large ones, while from MAE's perspective small errors have larger weight. To illustrate this, a small error of predicting 0.5 instead of 0 is taken as an example, which from the RMSE perspective, becomes a loss of 0.5^2, which is equal to a loss of 0.25. Whereas for the MAE, the loss is 0.5, thus greater than the RMSE. In this regard, the evaluation that one makes is that the errors that the plain regression model makes on the zeroes have a small distance from the real counts. Moreover, plain regression models make this error much more frequently than ZIR models. Hence, according to MAE, the ZIR models are incredibly better than the plain regression models. Although even through the RMSE, the ZIR models also show an improvement over the single regression models, even if the improvement is marginal in some cases.

In this study, one sought to build forecasting models to predict whether a bicycle would arrive at a given station in the next hour. Concomitantly with this goal, since the target variable is predominantly equal to zero, the reasoning followed in this study was devoted to understanding if models built within the Zero-Inflated Model framework are better at making these kinds of predictions. One can conclude that for the models developed in this study, there was a gain in prediction when developing such models in the Zero-Inflated framework. The models that were able to make the best predictions are the ZIR Random Forest and ZIR Gradient

Boosting based models, with an RMSE of 0.0669 and 0.0667 and an MAE of 0.0060 and 0.0061, respectively.

Still, it is recognized that there is much room for these predictions to be improved, such as hyper parameterization of the models and new feature engineering processes.

3.6. Deployment

As a deployment suggestion, the implementation of the contributions of this study is primarily aimed at improving the level of service delivery of bike-sharing systems at the user level. This means that the forecasts made, although in need of improvement, should be integrated into the application users use to check-in and check-out bicycles at the stations. Thus, if a user wants to start a trip and there is no bike available, the cyclist can check if any bike is expected to arrive in the next hour. With this integration, the value proposition of BSS would be enhanced, and since such an implementation increases the predictability of when a user would be able to start a trip, all the effort put into developing the model and implementing it is valuable. Some of the more traditional means of transportation, such as the bus, subway, and train, already provide ways to indicate and predict when a user will be able to continue their journey, so it is necessary to bring BSS closer to these more traditional modalities to contribute to the adoption of this sustainable transportation.

Monitoring and maintaining the deployed machine learning process is another very important aspect. In this study, the models were built considering only the North Hollywood bike stations, however, there are stations in other areas of the city, so a roll-out strategy needs to be developed to integrate other areas in the city into the current system. However, since the behavior patterns towards the stations are different, as new stations become part of the system, adjustments to the current model should be made. This is a continuous process, as political, economic, socio-cultural, and technological (PEST) aspects have a direct impact on how services operate and how people conduct their daily activities, changing previously pre-established patterns.

CHAPTER 4

Conclusions

4.1. Concluding remarks

One has argued throughout this work on the need to develop ways to improve the service provided by BSS from the end-user's perspective, filling a gap in the literature and providing an unprecedented methodology to improve BSS service. It was seen through the literature review that much of the knowledge produced is focused on service improvement more at the operational level, such as the bicycle rebalancing process. However, since perfect rebalancing still cannot be guaranteed across the entire BSS service, some bicycle stations will run out of bicycles available for a user to start their trip at some point.

In this sense, as some of the researchers (Demidova et al., 2022) pointed out, one of the impediments to greater adoption of bicycles as a common means of transportation in the daily life of an urban population is precisely because of the lack of predictability of availability. This study main goal was to the develop, from the cyclist’s perspective, a proof of concept on the feasibility of informing the user about the possibility of starting a trip in a pre-defined time interval by predicting whether bicycles will arrive at a given station. If affirmative, it was important to predict how many bicycles will be arriving.

Additionally, it was intended to analyze whether a Zero-Inflated Regression (ZIR) framework would be able to improve the forecast models developed. That said, one developed plain regression models based on Decision Trees, Random Forest and Gradient Boosting, and concomitantly, created new models based on these algorithms, but inserted them within the Zero-Inflated Model framework with the objective of understanding if their performance would improve when performing a classification task before regression. Besides these algorithms, Decision Tree with Ada Boost, Extra Trees and Gaussian Naive Bayes models were also developed in the Zero-Inflated framework.

The main contributions of this study are: (i) the ability to predict how many bicycles will arrive at a given station is a feasible improvement for BSS, as the models developed throughout this work are able to inform whether or not a bicycle will arrive at a station, thus proving that this concept is a viable functionality for BSS; (ii) the models developed through the Zero-Inflated Regression approach are a path that can be explored to improve prediction models

applied to the BSS when there is a high frequency of zero counts, once ZIR models outperformed the plain regressions models developed; (iii) unprecedented methodological contribution to the literature on BSS focusing on the end-user's decision power about whether or not it will soon be possible to start a trip. By proving that this concept is viable, this work provides a novel solution in the sense that since it is not possible to guarantee that there will always be bicycles available at a given station for a user to start a trip, it is important to provide the user with information whether or not it will be possible to use the BSS soon by predicting if bicycles will arrive at that station in the next few minutes.

Furthermore, this concept, of predicting whether a user will have bicycles available to start a trip soon, is applicable and generalizable to any fixed docks BSS where data is available on all trips arriving at each of the stations and there is a risk that there will be no bicycles available to start a trip. However, improving predictions by means of ZIR models are applicable only to systems where the bicycle count is zero-inflated, that is, the frequency of counting zeros is excessively higher than the other values.

4.2. Limitations and Future Research

It should be noted that this work focus was put on tree-based models' implementation. When looking at the performance of the models, the gain from fitting a tree model within the Zero-Inflated framework is modest but still can represent gains of around 10%, in terms of RMSE, for some of the tree-based algorithms discussed previously. Therefore, future research is needed to evaluate the behaviour of other algorithms in the ZIR framework and compare them with analogous plain regression models such as, among others, feedforward neural networks and support vector machines (SVM). During this dissertation, there was an attempt to apply the SVM machine learning technique, but due to the need to better investigate the hyperparametrization of the model, it was decided to consider other models instead of SVMs. Additionally, this work did not dedicate itself exhaustively to finding the best parameters for each of the models developed, so this is also a gap to be fulfilled.

The meteorological variables proved to be of great importance in this study; it is enough to observe the VIF they obtained before the construction of the forecast models to be clear that they undeniably play a notorious role in the forecasts. In this sense, it is beneficial to trigger specific studies with the aim of realizing that other exogenous variables can bring precious information to improve the performance of the forecast models.

Furthermore, incorporating findings from land use and origin-destination studies into the forecasting models of this work could also play a significant role in improving the performance

of the models, as they could help to understand fluctuations in the patterns of travel destinations. Therefore, there is plenty of room for new contributions in the BSS literature, regarding the interest in informing the user about how the situation in the stations will evolve. With this, there would be a greater chance of success for the development of bike station-based features.

Additionally, another variable that in the future could play a significant role in making predictions of how many bicycles would arrive at a given station is a feature that indicates an estimate based on transition probability, knowing the ongoing bicycle trips. Given all the ongoing trips that have not yet ended, what is the probability that this trip will end at the station where I want to predict how many bicycles will arrive?

Initially, this work created lags from the target variable to be used in the model, however, an analysis to identify the optimal number of lags would be necessary to offer improvements to the study carried out. In this sense, it was decided to build models without lag variables next, since with the arbitrary numbers from 1 to 6 lags, no improvement was identified.

Bibliographic References

Almannaa, M. H., Elhenawy, M., & Rakha, H. A. (2020). Dynamic linear models to predict bike availability in a bike sharing system. *International Journal of Sustainable Transportation*, *14*(3), 232–242. https://doi.org/10.1080/15568318.2019.1611976

Amalia, R. N., Sadik, K., & Notodiputro, K. A. (2021). A Study of ZIP and ZINB Regression Modeling for Count Data with Excess Zeros. *Journal of Physics: Conference Series*, *1863*(1). https://doi.org/10.1088/1742-6596/1863/1/012022

Ashqar, H. I., Elhenawy, M., Almannaa, M. H., Ghanem, A., Rakha, H. A., & House, L. (2017). Modeling bike availability in a bike-sharing system using machine learning. *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, 374–378. https://doi.org/10.1109/MTITS.2017.8005700

Ashqar, H. I., Elhenawy, M., Rakha, H. A., Almannaa, M., & House, L. (2021). Network and station-level bike-sharing system prediction: a San Francisco bay area case study. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. https://doi.org/10.1080/15472450.2021.1948412

Bacciu, D., Carta, A., Gnesi, S., & Semini, L. (2017). An experience in using machine learning for short-term predictions in smart transportation systems. *Journal of Logical and Algebraic Methods in Programming*, *87*, 52–66. https://doi.org/10.1016/j.jlamp.2016.11.002

Cavallaro, C., & Tramontana, E. (2021). User assistance for predicting the availability of bikes at bike stations. *CEUR Workshop Proceedings*, *2963*, 132–143.

Cenni, D., Collini, E., Nesi, P., Pantaleo, G., & Paoli, I. (2021). Long-term prediction of bikes availability on bike-sharing stations. *Proceedings - DMSVIVA 2021: 27th International DMS Conference on Visualization and Visual Languages*, 1–7. https://doi.org/10.18293/DMSVIVA2021-001

Cerqueira, S., Arsenio, E., & Henriques, R. (2021). On how to incorporate public sources of situational context in descriptive and predictive models of traffic data. *European Transport Research Review*, *13*(1). https://doi.org/10.1186/s12544-021-00519-w

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, B., Pinelli, F., Sinn, M., Botea, A., & Calabrese, F. (2013). Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 53–58. https://doi.org/10.1109/ITSC.2013.6728210

Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., Wu, Z., Pan, G., Nguyen, T.-M.-T., & Jakubowicz, J. (2016). Dynamic cluster-based over-demand prediction in bike sharing systems. *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 841–852. https://doi.org/10.1145/2971648.2971652

Chen, Z., Wu, H., O'Connor, N. E., & Liu, M. (2021). A Comparative Study of Using Spatial-Temporal Graph Convolutional Networks for Predicting Availability in Bike Sharing Schemes. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2021-Septe*, 1299–1305. https://doi.org/10.1109/ITSC48978.2021.9564831

Choudhary, G., & Narayan Singh, S. (2020). Prediction of heart disease using machine learning algorithms. *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, 197–202. https://doi.org/10.1109/ICSTCEE49637.2020.9276802

Demidova, N., Novakovic, A., & Marshall, A. H. (2022). Optimizing the Belfast Bike Sharing Scheme. In *Lecture Notes in Networks and Systems* (Vol. 295). https://doi.org/10.1007/978-3-030-82196-8_43

Ethridge, J., Ditzler, G., & Polikar, R. (2010). Optimal ν-SVM parameter estimation using multi objective evolutionary algorithms. *2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 IEEE Congress on Evolutionary Computation, CEC 2010*. https://doi.org/10.1109/CEC.2010.5586029

Feng, C., Hillston, J., & Reijsbergen, D. (2017). Moment-based availability prediction for bike-sharing systems. *Performance Evaluation*, *117*, 58–74. https://doi.org/10.1016/j.peva.2017.09.004

Gast, N., Massonnet, G., Reijsbergen, D., & Tribastone, M. (2015). Probabilistic forecasts of bike-sharing systems for journey planning. *International Conference on Information and Knowledge Management, Proceedings*, *19-23-Oct-*, 703–712. https://doi.org/10.1145/2806416.2806569

Ge, Y., Qu, W., Qi, H., Cui, X., & Sun, X. (2020). Why people like using bikesharing: Factors influencing bikeshare use in a Chinese sample. *Transportation Research Part D: Transport and Environment*, *87*. https://doi.org/10.1016/j.trd.2020.102520

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Jiang, J., Lin, F., Fan, J., Lv, H., & Wu, J. (2019). A Destination Prediction Network Based on Spatiotemporal Data for Bike-Sharing. *Complexity*, *2019*. https://doi.org/10.1155/2019/7643905

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality and Quantity*, *52*(4), 1957–1976. https://doi.org/10.1007/s11135-017-0584-6

Jun, M.-J. (2021). A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area. *International Journal of Geographical Information Science*, *35*(11), 2149–2167. https://doi.org/10.1080/13658816.2021.1887490

Kambezidis, H. D. (2012). The solar resource. *Comprehensive Renewable Energy*, *3*, 27–84. https://doi.org/10.1016/B978-0-08-087872-0.00302-4

Kim, M., & Cho, G.-H. (2021). Analysis on bike-share ridership for origin-destination pairs: Effects of public transit route characteristics and land-use patterns. *Journal of Transport Geography*, *93*. https://doi.org/10.1016/j.jtrangeo.2021.103047

Lam, A., Schofield, M., & Shyang Ho, S. (2019). Detecting (unusual) events in urban areas using bike-sharing data. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Local Events and News, LENS 2019*. https://doi.org/10.1145/3356473.3365190

Liu, X., Gherbi, A., Li, W., & Cheriet, M. (2019). Multi features and multi-time steps LSTM based methodology for bike sharing availability prediction. *Procedia Computer Science*, *155*, 394–401. https://doi.org/10.1016/j.procs.2019.08.055

Liu, X., & Pelechrinis, K. (2021). Excess demand prediction for bike sharing systems. *PLoS ONE*, *16*(6 June). https://doi.org/10.1371/journal.pone.0252894

Lucas, V., & Andrade, A. R. (2021). Predicting hourly origin–destination demand in bike sharing systems using hurdle models: Lisbon case study. *Case Studies on Transport Policy*. https://doi.org/10.1016/j.cstp.2021.10.003

Macioszek, E., Świerk, P., & Kurek, A. (2020). The bike-sharing system as an element of enhancing sustainable mobility - A case study based on a city in Poland. *Sustainability (Switzerland)*, *12*(8). https://doi.org/10.3390/SU12083285

Metro Bike Share. (2022). *Data*. https://bikeshare.metro.net/about/data/

Miyazawa, S., Song, X., Xia, T., Shibasaki, R., & Kaneda, H. (2019). Integrating GPS trajectory and topics from Twitter stream for human mobility estimation. *Frontiers of Computer Science*, *13*(3), 460–470. https://doi.org/10.1007/s11704-017-6464-3

Ruffieux, S., Mugellini, E., & Abou Khaled, O. (2019). Predictive Modeling for Optimization of Field Operations in Bike-Sharing Systems. *Proceedings - 6th Swiss Conference on Data Science, SDS 2019*, 119–124. https://doi.org/10.1109/SDS.2019.00011

Salih-Elamin, R., & Al-Deek, H. (2020). Short-term prediction for bike share systems' travel time under the effects of weather conditions. *Advances in Transportation Studies*, *50*, 81–94. https://doi.org/10.4399/97888255317326

Sardinha, C., Finamore, A. C., & Henriques, R. (2021). *Context-aware demand prediction in bike sharing systems: incorporating spatial, meteorological and calendrical context; Context-aware demand prediction in bike sharing systems: incorporating spatial, meteorological and calendrical context*. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, 353–366. https://doi.org/10.1016/j.comcom.2020.02.007

Soheil, S., Paleti, R., Balan, L., & Cetin, M. (2020). Real-time prediction of public bike sharing system demand using generalized extreme value count model. *Transportation Research Part A: Policy and Practice*, *133*, 325–336. https://doi.org/10.1016/j.tra.2020.02.001

Thu, N. T. H., Thanh, L. T., Dung, C. T. P., Linh-Trung, N., & Le, H. V. (2017). Multi-source data analysis for bike sharing systems. *International Conference on Advanced Technologies for Communications*, *2017-Octob*, 235–240. https://doi.org/10.1109/ATC.2017.8167624

Wang, J., Li, F., Yang, S., Li, Y., & Wang, Y. (2022). A Real-Time Bike Trip Planning Policy With Self-Organizing Bike Redistribution. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10646–10661. https://doi.org/10.1109/TITS.2021.3095177

Wang, K., Akar, G., & Chen, Y.-J. (2018). Bike sharing differences among Millennials, Gen Xers, and Baby Boomers: Lessons learnt from New York City's bike share. *Transportation Research Part A: Policy and Practice*, *116*, 1–14. https://doi.org/10.1016/j.tra.2018.06.001

Xu, C., & Wang, C. (2019). Analysis of E-bike Trip Duration and Frequency by Bayesian Duration and Zero-inflated Count Models. *KSCE Journal of Civil Engineering*, *23*(4), 1806–1818. https://doi.org/10.1007/s12205-019-0674-1

Yoshida, A., Yatsushiro, Y., Hata, N., Higurashi, T., Tateiwa, N., Wakamatsu, T., Tanaka, A., Nagamatsu, K., & Fujisawa, K. (2019). Practical End-to-End Repositioning Algorithm

for Managing Bike-Sharing System. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 1251–1258. https://doi.org/10.1109/BigData47090.2019.9005986

Zhang, C., Zhang, L., Liu, Y., & Yang, X. (2018). Short-term Prediction of Bike-sharing Usage Considering Public Transport: A LSTM Approach. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-Novem*, 1564–1571. https://doi.org/10.1109/ITSC.2018.8569726

Zhang, H. (2004). The Optimality of Naive Bayes. *FLAIRS Conference*.