A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

MASSIMO PIO CAROLLO

61400

A study on variability of cost difference between Green and Yellow NYC taxi rides over time

As a part of the Work Project "Urban Transportation in New York City: A Comparative Analysis of Accessibility, Pricing, and Usage Trends" with ANGELINA SUCHKOVA, WALE DOURA, PARAM SUMIRAN, LAURENZ VON PERBANDT

16/12/2024

Abstract

This thesis examines the interplay across diverse mobility options in New York City, focusing on cost variability and commuter behavior. This work investigates the evolution of fare disparities between taxis, using comprehensive datasets from NYC's Taxi and Limousine Commission, along with advanced predictive modeling approaches such as Long Short-Term Memory and Gradient Boosting. The study revealed how commuter choices are influenced by operational regions and pricing models: key findings highlight significant temporal and spatial fare trends. Enhanced taxi coverage in NYC's outer boroughs has led to a measurable shift in subway ridership. Green Taxis, designed to address transportation gaps in underserved neighborhoods, exhibit limited success, particularly in low-income areas, where reliance on private vehicles remains dominant. Additionally, a comparative analysis of dynamic versus fixed pricing models showcased the flexibility of ride-hailing platforms like Uber in addressing peak demands and spatial variations, differing from the rigidity observed in traditional pricing systems. These analyses provide critical insights into spatiotemporal variations in taxi demand, including the differential impacts of public transit availability and demographic factors.

By addressing pricing disparities and enhancing service accessibility, this work concludes by presenting actionable recommendations to improve equity and efficiency within NYC's transportation network. The findings aim to guide policymakers in developing adaptive and sustainable mobility strategies tailored to the evolving urban landscape.

Table of Contents

For the group part with ANGELINA SUCHKOVA, WALE DOURA, PARAM SUMIRAN, LAURENZ VON PERBANDT MASSIMO PIO CAROLLO

- **Introduction 5**
- **Literature Review 8**
- **NYC Transportation Industry Overview 17**

3.1 The historical development of NYC's transportation system 17

3.2 Key Boroughs of New York City: Economic and Cultural Centers 20

3.3 Pricing models and Regulations 22

- **4. Methodology and Findings 25**
- **5. Introduction to Individual Parts 28**

The individual part:

- **8. By Massimo Pio Carollo: A study on variability of cost difference between Green and Yellow NYC taxi rides over time.....30**

8.1 Introduction 30

8.2 Literature Review..... 31

8.3 Methodology..... 34

8.4 Models: Long Short Term Memory and Gradient boost model 38

8.5 Conclusions 44

13. Resources 46

Tables of figures

Figure 1. The correlational matrix of the matching rides dataset..... 37

Figure 2. Long Short-Term Memory Hourly model predicted results for Cost difference 40

Figure 3. Long Short-Term Memory Daily model predicted results for Cost Difference..... 41

Figure 4. Visualization of Gradient Boosting prediction model of Cost_difference 44

Figure 5. Scatter plot of predicted values versus actual values for a model predicting Cost Difference (Cost_difference) using Gradient Boosting..... 44

- **Introduction**

Urban mobility systems are critical to the functioning of modern cities, enabling the movement of millions of individuals daily while shaping economic, social, and environmental outcomes. In New York City (NYC), a city renowned for its dense population and complex infrastructure, transportation services play a pivotal role in supporting urban life. Over the years, NYC's transportation landscape has evolved significantly, reflecting changes in commuter needs, technological advancements, and policy-driven reforms. Central to this system are its taxi services, which range from traditional yellow and green taxis to app-based ride-hailing platforms like Uber and Lyft, and the broader network of public transit systems such as buses and subways.

Despite this diversity, challenges such as geographic inequities, congestion, and fluctuating demand persist, necessitating a deeper understanding of how different mobility modes interact and adapt within the city's unique urban environment. Efforts to expand and diversify NYC's transportation network have resulted in significant changes to its taxi ecosystem. Historically dominated by yellow taxis, the introduction of green taxis sought to extend services to underserved areas like Brooklyn, Queens, and the Bronx. These changes raised questions about how increased taxi coverage has influenced commuter behavior in the outer boroughs. Has this initiative successfully addressed gaps in accessibility, or has it merely shifted existing dynamics? Understanding these spatial patterns offers insights into the effectiveness of such policy-driven expansions.

At the same time, the interplay between public transit systems and taxi services highlights the interconnectedness of NYC's mobility ecosystem. Public buses, often serving as a lifeline for disadvantaged communities, experience competition and complementarities with taxis, especially in areas where subway access is limited. Changes in bus service availability, delays, or coverage

can influence taxi usage, creating ripple effects across the transportation network. Exploring these dynamics sheds light on how public and private modes of transport coexist and compete within a constrained urban space.

Pricing models further complicate NYC's mobility landscape. The cost differences between yellow and green taxis provide a lens to examine how operational areas and fare structures influence commuter choices. These variations reflect not only economic factors but also temporal and spatial nuances, revealing how fares align—or fail to align—with demand patterns. While green taxis aim to serve peripheral neighborhoods, the cost dynamics between the two services prompt questions about fairness, efficiency, and accessibility.

The emergence of app-based ride-hailing services such as Uber and Lyft has introduced new layers of complexity to NYC's transportation system. These platforms, offering flexibility and convenience, cater to a wide range of travel purposes, from routine commutes to leisure activities. By comparing trip patterns between green taxis and ride-hailing services, it is possible to uncover how user preferences and behaviors diverge across different modes. These insights are critical for understanding the broader implications of these services on traditional taxi operations and commuter mobility.

Underlying many of these discussions is the tension between fixed and dynamic pricing models. Traditional taxis operate under regulated pricing systems, offering predictability but limited adaptability to changing conditions. In contrast, ride-hailing platforms leverage dynamic pricing algorithms to adjust fares in real time based on demand and traffic. These distinct approaches offer a unique opportunity to explore how pricing strategies influence fare sensitivity and commuter decision-making, especially in a city as multifaceted as NYC.

This is why this thesis investigates several interconnected research angles of New York City's transportation systems: by examining the evolution of fare disparities between yellow and green taxis, along with employing advanced predictive models to understand cost variations over time and space, it was possible to show high variability in cost differences between comparable taxi rides and the consequent difficulty in predicting those values. Additionally, green taxis primarily serve outer boroughs with stable fares, while yellow taxis dominate central Manhattan, often with higher and more variable costs during peak hours and weekends, highlighting how operational zones and service shape commuter behavior.

Moreover, this study reveals how shifts in taxi coverage influence subway ridership in underserved areas of NYC. Another aspect involves evaluating the impact of dynamic pricing models as seen in ride-hailing platforms like Uber and contrasting them with traditional systems used by taxis. This led to a deep understanding of flexibility and efficiency in addressing peak demand and spatial variability.

- **Literature Review**

2.1 Foundations of Urban Mobility

Urban mobility describes the systems, services, and infrastructure that enable the movement of people and goods within urban areas. It is defined as the integration of multiple transportation modes, such as public transit, private vehicles, shared mobility, walking, and cycling, and it aims to address challenges of accessibility, sustainability, and equity. Its scope comprises multimodal systems for seamless travel, sustainability efforts to reduce environmental impact, and the adoption of advanced technologies like AI and ICT to improve efficiency and user experiences. An important aspect of urban mobility is to ensure economic and social equity, making transportation affordable and accessible for all residents while balancing environmental and economic goals (Arthur D. Little, 2011). The evolution of urban transportation systems was influenced by sociopolitical and historical developments. In pre-20th-century cities, compact and walkable designs dominated, with limited mechanized transport. The post-war era saw a surge in car ownership, resulting in investments in road networks and the neglect of tram systems in many urban areas. However, some cities, such as Karlsruhe, maintained and innovated tram networks, in contrast to car-centric developments in other cities. The late 20th century saw a rebirth of interest in sustainable transit, which led to innovations like tram-train systems and integrated urban planning (Pflieger, 2009). To this day, private cars remain dominant globally, accounting for 45% of trips. However, projections suggest a 15% decline by 2035, driven by the growth of autonomous vehicles and micromobility (McKinsey & Company, 2023). Urban transportation systems show path dependencies, where historical decisions, such as the introduction or removal of tram systems or the growth of private car ownership, create lasting influences on urban policies and infrastructure. The high cost of reversal locks cities into specific development trajectories unless

significant crises or innovations occur (Pflieger, 2009). New York City shows unique urban mobility characteristics shaped by its geography and central transportation hubs. Sustainable transportation modes, such as walking, biking, and public transit, account for more than two-thirds of all trips. NYC's dense urban cores, particularly in Manhattan and key business districts, play a critical role in that. The Central Business District (CBD) experiences heavy congestion, with sustainable transportation accounting for 78% of trips into the area. Public transit remains vital for most residents, although aging infrastructure and telecommuting have caused slight declines in subway use (New York City Department of Transportation, 2019). Programs like Citi Bike illustrate the emergence of micromobility, reshaping short-distance travel in areas like Midtown Manhattan (Sun & Axhausen, 2016).

NYC's mobility patterns reflect interconnected yet constrained boroughs, with bridges and tunnels serving as critical, often congested nodes. The growth of freight and e-commerce is adding pressure to urban traffic systems. Geographical constraints, such as the city's island structure, limit road expansion and shift the focus to efficient public and shared transportation. Additionally, high tourist volumes, especially in areas like Times Square, require special mobility solutions for localized travel (New York City Department of Transportation, 2019).

2.2 Role of Taxis in NYC's Urban Mobility

Taxis have been a significant part of New York City's transportation landscape for decades. They serve as a key alternative to public transit, particularly in areas with limited-service coverage, and for seniors and individuals with mobility challenges (New York City Department of City Planning, 2009). Historically, NYC's taxi system was governed by the medallion system, which regulated the number of cabs operating within the city. However, the rise of app-based ride-hailing services,

such as Uber and Lyft, has caused a significant decline in traditional yellow taxi usage, reshaping urban mobility patterns and changing urban transportation planning. The transition from medallion taxis to ride-hailing services indicates a broader shift in consumer preferences, technological adoption, and the evolving role of taxis in the city's transportation structure (Moro, 2021; New York City Department of City Planning, 2009).

This changing landscape has significant implications for urban planning and mobility strategy in NYC. As ride-hailing services increasingly dominate, the city must address issues of equity, congestion, and sustainability to ensure that taxis, both traditional and app-based, continue to contribute effectively to the broader transportation network (Moro, 2021).

2.3 Socio-Economic and Demographic Factors Influencing Urban Mobility in NYC

Income inequality significantly impacts access to transportation options in New York City, showing mobility inequalities among different socio-economic groups. Lower-income individuals face barriers, such as high taxi costs and inconsistent public transport availability in underserved neighborhoods, limiting their ability to access essential services like jobs and healthcare (New York City Department of City Planning, 2009). In contrast, wealthier residents benefit from better access to diverse transportation modes, including taxis and ride-hailing services, and are more likely to explore geographically diverse areas, reducing experienced segregation (Moro, 2021). High-income areas, particularly in central Manhattan, are dominated by yellow taxis, while peripheral and low-income areas often rely on less reliable livery cabs or limited public transit options, deepening mobility inequalities (New York City Department of City Planning, 2009). This dynamic reinforces systemic segregation and unequal access to opportunities, as residents in

these areas face constraints tied to economic and infrastructural limitations (Moro, 2021). Work-from-home (WFH) and hybrid work models have also transformed commuting patterns in the city. While these models have reduced rush-hour congestion, they have increased off-peak traffic, particularly in areas with flexible work environments. This shift required adaptations in transit schedules and infrastructure to adapt to the changing traffic flows (Lasley, 2021).

2.4 Public Policy and Governance in Urban Mobility

By aiming to address social equity, environmental sustainability, and technological advancements, public policy and governance play a critical role in shaping urban mobility systems. While stable political regimes rather reinforce existing policies, institutional and political changes either strengthen innovation or serve as barriers to reform. Urban transport policies, such as integrated planning in cities like Grenoble, demonstrate how aligning transportation with urban development can result in significant social and spatial benefits, including reduced car reliance and improved accessibility (Pflieger,2009).

Urban and transport planning also have direct health implications. Proper planning can mitigate key urban exposures, such as air pollution, noise, and urban heat islands, while enhancing access to green spaces and opportunities for physical activity (Nieuwenhuijsen, 2016). Community-level interventions, such as promoting green spaces and reducing car reliance, have proven more cost-effective and impactful than individual-level efforts. These strategies emphasize equitable access to transport systems, bridging the gap between income-segregated communities by expanding affordable public transit and reducing cost barriers for taxis (Moro, 2021). In New York City, regulation is a key tool for managing mobility challenges. The medallion system historically

controlled the taxi industry, while congestion pricing programs are now being introduced to reduce vehicle volumes in the Central Business District (Miskolczi, 2021; New York City Department of Transportation, 2019). Cities worldwide are adopting similar measures, including parking restrictions, car-free zones, and dynamic tolling, to manage traffic and pollution. Such strategies are crucial in balancing technological advancements with societal acceptance and legal frameworks, especially for innovations like autonomous vehicles (Bouton,2015). Public-private partnerships (PPPs) further address urban mobility challenges by fostering collaboration between governments, private companies, and tech innovators. These partnerships accelerate the deployment of solutions like electric vehicle (EV) charging stations, shared mobility platforms, and autonomous vehicle (AV) pilot projects. For example, governments can subsidize shared mobility services, provide tax benefits for EV adoption, and promote subscription-based services to encourage behavioral shifts toward sustainable mobility (Miskolczi,2021;Butler,2020;Kamargianni,2016).

Finally, governance frameworks must address data privacy and transparency, ensuring user trust while leveraging mobility data for planning and optimization. Policies supporting AV integration, such as liability guidelines and safety standards, can build public confidence and accelerate adoption (Miskolczi, 2021). By investing in intelligent transportation systems (ITS), high-speed internet, and secure communication networks, cities can establish the foundations for smart infrastructure and future-ready urban mobility systems (Butler, 2020).

2.5 Technological Innovations and Smart Mobility in NYC

Technological innovation in urban mobility is expected to play an important role in dealing with the growing transportation challenges. Over-reliance on private vehicles and outdated

infrastructure is pushing many urban mobility systems toward breakdown. To address these challenges, advancements in automation, shared mobility, and electrification are transforming urban transportation systems (Miskolczi, 2021). Services like Uber and Lyft have revolutionized urban transit by introducing e-hailing, car sharing, and on-demand shuttles. These innovations are driving a decline in car ownership, especially among younger generations in developed nations, as preferences shift toward shared mobility. Mobility-as-a-Service (MaaS) platforms exemplify this transformation by integrating public transit, car sharing, and bike-sharing into a single, seamless system (Bouton, 2015). Through real-time data sharing and journey planning, MaaS offers door-to-door mobility while reducing reliance on personal vehicles (Butler, 2020). However, its implementation faces challenges, such as limited integration across operators, complex revenue-sharing models, and varied user willingness to pay for subscription-based services (Kamargianni, 2016). Shared mobility services, including dockless bike-sharing programs and app-based ridesourcing, address critical issues like first- and last-mile connectivity. They also deliver environmental benefits by reducing vehicle miles traveled and decreasing road congestion. These innovations are key to strengthen multimodal travel, sustainability, and urban accessibility (Butler, 2020). Emerging technologies like big data and the Internet of Things (IoT) are central to modern urban traffic management. Inefficiencies in urban mobility currently cost cities 2–4% of their GDP due to wasted time and resources (Bouton, 2015). Intelligent Transportation Systems (ITS) integrate advanced ICT and real-time analytics to optimize transportation networks. Applications include adaptive traffic signal control, incident detection, and dynamic scheduling for public transit. Additionally, networked ecosystems, such as vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) systems, promise enhanced safety and traffic flow (Lasley, 2023; Moss, 2012; Butler, 2020).

The transition to electric and autonomous vehicles (AVs) is redefining urban mobility. Electric vehicles (EVs) significantly improve energy efficiency and reduce emissions (Moss, 2012; Bouton, 2015). Meanwhile, AVs, capable of navigating without human input, hold the potential to reduce road accidents by up to 90%, increase traffic efficiency, and provide mobility options for individuals without driving licenses. However, challenges, such as cybersecurity risks, ethical dilemmas during accidents, and infrastructure readiness, remain significant barriers to widespread adoption (Butler, 2020). Adoption might remain slow due to economic and social barriers, which highlights the need for regulatory clarity and collaborative efforts across sectors (Bouton, 2015). Crucial for supporting these innovations are investments in smart infrastructure, including electric vehicle charging networks and data processing for traffic management (Moss, 2012).

2.7 Comparative Studies and Lessons from Other Cities

New York City may benefit from examining sustainable and innovative mobility practices implemented in other global cities. One significant area of innovation is the development of Mobility-as-a-Service (MaaS) platforms. Cities like Helsinki and Gothenburg have pioneered MaaS initiatives, such as UbiGo, which integrate various transit modes into a single, seamless system. These platforms allow users to plan, book, and pay for multi-modal journeys across different transport providers, simplifying access to sustainable transport and reducing reliance on personal vehicles (Bouton, 2015; Kamargianni, 2016). Other examples of successful MaaS implementations include the Octopus card in Hong Kong and the Oyster card in London. These smart card systems have significantly increased public transit usage by offering a unified payment method for various modes of transportation, including buses, subways, and ferries. The convenience and efficiency provided by these integrated systems highlight the potential for MaaS

to transform urban mobility by promoting multimodal travel and making public transportation more user-friendly (Butler, 2020). However, transferring global best practices to New York City's context requires careful consideration of the city's unique characteristics. Different city types—such as megacities, rising cities, and car-dominated cities—experience unique mobility evolutions based on factors like density, existing infrastructure, and available resources. For instance, while NYC shares similarities with other megacities in terms of scale and complexity, its legacy infrastructure and regulatory environment may present challenges not encountered in cities like Helsinki or London (Bouton, 2015).

2.8 Emerging Trends and Future Challenges in Urban Mobility

Urban mobility is experiencing significant transformations driven by emerging trends and future challenges. By 2030, 60% of the global population will live in cities, intensifying pressure on urban infrastructure (Moss, 2012; Bouton, 2015). Integrating urban mobility with land-use planning is becoming essential for future development, ensuring that transportation systems align with urban growth strategies and environmental objectives (Bouton, 2015). There is a growing reliance on mixed transportation modes, including public transit, car-sharing, biking, and walking. This shift indicates urban residents' increasing preference for connectivity and access over private car ownership (Moss, 2012). Reallocating urban space is another emerging trend aimed at enhancing the quality of urban life. This involves repurposing areas previously dedicated to parking for green spaces, bike-sharing docks, or electric vehicle (EV) charging stations. Introducing car-free zones and low-emission areas helps reduce congestion and pollution, improving urban quality of life (Moss, 2012; Moskolczi, 2021).

The traffic composition is also evolving. Congestion has increased due to the surge in e-commerce demand, leading to significant delays during peak periods, especially in densely populated urban areas (Lasley,2023).

Key challenges include the slow adoption of innovations due to economic and social barriers, the need for integrated infrastructure to support shared and autonomous systems, and balancing technological advancements with societal acceptance and legal frameworks (Miskolczi, 2021). These emerging trends and challenges highlight the necessity for cities like New York to adapt and innovate. Embracing integrated planning, promoting multimodal transportation options, and preparing for future demographic shifts are crucial steps toward creating a resilient and efficient urban mobility system.

- **NYC Transportation Industry Overview**

3.1 The historical development of NYC's transportation system

The transportation system of New York City lays the foundation of the world's most dynamic city, reflecting centuries of technological development, policy experimentation, and adaptation. The system comprises ferries, buses, subways, and taxis and regularly makes more than 8.5 million trips each day to meet the needs of New York City's growing population, tourism, and economy. We're going to trace the development of NYC's transportation system especially buses, subways, and taxis, providing the historical context needed to understand the interconnections and evolution of urban mobility in New York City.

Buses: From Horsepower to High-Tech

Beginning with horse-drawn omnibuses, NYC's bus system developed in the early 19th century. These basic buses traveled along fixed routes and charged fares of 12.5 cents, or $4.50 today. By the 1850s, omnibuses were being replaced by horse-drawn streetcars, allowing for more speed and capacity. Electric trolleys appeared in the 1880s, allowing a new era of mechanized transit. Motorized buses hit New York City streets in 1907 with the Fifth Avenue Coach Company, offering a cleaner, more efficient alternative to horse-drawn systems. The city purchased private bus lines during the Great Depression and brought all operations under public control by the 1940s. By 1950, buses were carrying 2.3 million daily passengers, connecting neighborhoods often not served by the subway. Select Bus Service (SBS), introduced in 2008, modernized the bus network and reduced travel times by 20–30% with features such as off-board fare collection and dedicated lanes. By 2020, NY's buses served 1.2 million daily riders, down from their mid-20th-century peak but still important for underprivileged areas. Recent efforts have turned to electrification; in 2022, the MTA announced that it planned to switch its entire fleet to zero-emission buses by 2040.

Subways: The Backbone of Urban Transit

The subway system, one of the most important components for public transportation, originates back to the overcrowded streets of the 19th century, where the trains or “El trains” provided the first solution. The first subway line, run by the Interborough Rapid Transit Company (IRT), opened on Oct. 27, 1904. Its first day of service had more than 150,000 passengers, running from City Hall to 145th Street. The fare was fixed at five cents, an affordable method of transportation for the city’s working class. In the early years, the IRT and other systems that emerged together to become known as the Brooklyn-Manhattan Transit (BMT) system funded a massive expansion dubbed the Dual Contracts, doubling the size of the network by 1913. In 1940, the IRT, BMT, and the Independent Subway System (IND) were unified under public control, simplifying fares and operations, which resulted in 2 billion trips annually by the 1950s.
The ’70s and ’80s were dark decades for the subway. Underfunding, poor maintenance, and rising crime impacted ridership to record the lowest rates, which made the M.T.A. launch a $54 billion capital program to improve infrastructure by 1981, including replacing the old tracks and rolling stock. In 1993, the MetroCard was launched, a switch to modernization for both subway and bus systems, and in 2000, more major investments were made to gain public trust, which led us to today. Comprising 472 stations, the subway serves about 5 million weekday riders, about 68% of New Yorkers. Its ongoing rollout of OMNY contactless payments and accessibility upgrades is proof of its commitment to modernizing its network. However, challenges still exist, including old infrastructure, delays, and declining ridership in a post-COVID world that still needs improvements and consistent investments.

Taxis: From Hackney Carriages to Uber Dominance

New York City's taxi industry has been an essential component of the city's transportation system since the mid-19th century, when horse-drawn hackney carriages first traveled the cobblestone streets. These early carriages were used as a luxurious service on demand for wealthier residents by charging passengers for rides between locations. The New York Taxicab Company appointed gasoline taxis in 1907, with 600 painted bright red and green. Motorized cabs gained popularity because they were a faster and more reliable option in comparison with horse-drawn alternatives. By 1910, over 1,000 gasoline-powered cabs flooded the streets of NYC, paving the way for the modernized industry seen today. In the 1920s, Hertz Rent-A-Car founder John Hertz debuted the iconic yellow taxi that we know today. The yellow color was selected because a 1960s study concluded it was the most visible color from a far distance, allowing passengers to identify taxis easily in the thriving streets of Manhattan. By the 1930s, taxis were a familiar sight on New York City's streets, providing cheap, convenient rides for a growing population. The medallion system was introduced in 1937 to regulate how many cabs could operate in the streets and to guarantee a certain level of service. The system limited the number of medallions to 13,595, a figure that remained constant for decades. Although this policy increased accountability and system stability, it blocked potential players, new drivers, and organizations from entering the industry. By the mid-20th century, medallion ownership had become a significant investment with values surpassing $1 million. By the early 2010s, Uber and other ride-hailing services were upending the taxi industry, shaking its dominance and reshaping the transportation market. Uber's app-based platform debuted in 2011 with upfront prices, shorter wait times, and convenient fares. Within a year, ride-hailing services began to appear, with Lyft also taking a significant market share. In 2018, NYC advanced a cap on growth, aiming to decrease the number of ride-hailing vehicles on the streets to stop oversaturation. The city launched congestion pricing zones in Manhattan to ease traffic flow

and balance the competition between taxis and ride-hailing services. But the medallion system, which had always been stable and served as a symbol of the taxi industry, crashed. Medallion values hit rock bottom, leaving many owners bankrupt and their investments worthless. By 2019, the effect of ride-hailing was indisputable. In NYC alone, Uber recorded 91 million trips, while yellow cabs took just 77 million trips. In recent years, the focus has shifted to sustainability and accessibility. New York City, in 2022, started a pilot program to replace older taxis with electric models, aiming to electrify a third of the taxis by 2030. Still, challenges remain since the current system doesn't prioritize the needs of wheelchair-accessible vehicles (WAVs) and disabled passengers. While yellow cabs are an iconic symbol of New York City, their continued relevance in the face of changing consumer preferences and technology is a question mark. Ensuring that taxis are not just a prominent mode in the city's multimodal transportation future will require further digitized platforms, sustainable vehicles, and regulations focused on equity.

From subways breaking down barriers of surface congestion and enabling mass transit easily across the city, buses improving access to underserved neighborhoods, and taxis providing on-demand service, each mode of transportation has had a unique impact on the city's growth, boosting tourism and supporting the economy. They together represent not only technological progression but also NYC's capacity to adjust to the needs of an ever-evolving city. This evolution allows us to see in what measure different transportation modes influence each other.

3.2 Key Boroughs of New York City: Economic and Cultural Centers

The city's transportation system is thus a lifeline for its residents. There are several high-density populated areas located in New York, which serve as the focal point for a lot of the economic and cultural activities of the city. Areas such as Downtown Brooklyn,

Flushing Queens, Midtown Manhattan, and Bronx Grand Concourse are some of these key areas, which cluster a lot of the city's economic and cultural activities. Let us investigate some of the key characteristics of the 5 boroughs of New York City.

Manhattan: It is the financial and cultural heart of the city. During the day, it has the highest population density in the world because of its accumulation of businesses and cultural attractions. Iconic neighborhoods like Midtown Manhattan and the Financial District are served by key transportation hubs such as Grand Terminal Station and the Port Authority Bus Terminal.

Brooklyn: Once a center for docks and factories, it has rapidly gentrified at the turn of this century into one of New York's most iconic and influential areas. Between 2010 and 2015, it has seen the greatest number of jobs being added compared to any of the boroughs outside Manhattan, adding 106,000 jobs. Its location serves to be a key area for the city's transportation network, as it connects the residential areas of Long Island with Manhattan via bridges, tunnels, and the subway.

Queens: It is the most ethnically diverse borough of the city, having some of the highest numbers of foreign-born residents, accounting for 47% of its population. It serves as a major residential zone in the city, having key population clusters such as Flushing and Jamaica. It also hosts the two main airports of New York: La Guardia and JFK Airport.

Bronx: It is a traditionally underserved region of New York, largely being a residential hub for the relatively poor population of the city. It is located very close to Manhattan, thereby making it a key source of labor for the city's economy. It also faces some of the highest commuting times of any of the boroughs in the city. The Metro-North commuter rail serves as the major transportation lifeline for its residents.

Staten Island: It is the least densely populated area of the city, characterized by its suburban setting and relative disconnection from the rest of the city. Often referred to as the 'forgotten

borough' of the city, it has the highest rates of car ownership in the city, which is highlighted by the limited public transportation options available. The Staten Island Ferry serves as a major link to its residents, connecting them to Manhattan, whereas the Verrazano-Narrows Bridge provides the crucial roadway linkage to Brooklyn.

3.3 Pricing models and Regulations

Yellow Taxis and the Medallion System

To ensure accuracy, historically, taxi fares were calculated using mechanical or electronic taximeters calibrated periodically, with prices based on time and distance. Consequently, operational costs—fuel prices, maintenance, and inflation—led to occasional fare adjustments. This is why fixed meter rates were often revised annually or in response to economic changes. This system was created to control the number of taxis on city streets, ensuring service quality, driver income stability, and market predictability. The medallion became a tradable commodity, its value fluctuating with market demand and TLC policies. Yellow taxis primarily operate in Manhattan and nearby areas, where street-hail demand is highest. They are equipped with meters that calculate fares based on time and distance, with additional surcharges for specific conditions:

- **Peak Hours:** An extra fee applies during weekday rush hours.
- **Congestion Surcharge:** A flat fee of $2.50 is levied for trips below 96th Street in Manhattan, introduced in 2019 to alleviate traffic and support public transit funding.

Despite their dominance, yellow taxis have faced criticism for limited coverage outside central Manhattan and for their rigidity in responding to real-time fluctuations in demand.

Introduction of Green Taxis (Boro Taxis)

In response to the geographic service disparity, TLC introduced green taxis, also known as Street Hail Liveries (SHLs), in 2013. These vehicles were designed to address the transportation needs of NYC's underserved outer boroughs (Brooklyn, Queens, Staten Island, and the Bronx) and northern Manhattan. Unlike yellow taxis, green taxis are prohibited from picking up street hails in Manhattan's central business district and at city airports, except through pre-arranged rides. Green taxis operate under the same regulated pricing model as yellow taxis, with fares determined by meters based on time and distance. The introduction of green taxis marked an important step in addressing the transportation inequities in NYC, providing more accessible and affordable options for residents in low-demand areas.

Operational and Regulatory Distinctions

The distinction between yellow and green taxis extends beyond geography to operational and market dynamics:

- **Street Hail Zones:** Yellow taxis have exclusive rights to street hails in central Manhattan, while green taxis serve outer boroughs.
- **Meter-Based Pricing:** Both services share the same regulated pricing structure, but green taxis rarely encounter high-density zones with congestion surcharges.
- **Licensing and Oversight:** Both are regulated by the TLC, ensuring driver and passenger safety, service quality, and fare transparency.

While both services follow traditional meter-based pricing, the geographic segmentation reflects an attempt to balance demand across the city, mitigating the dominance of yellow taxis in high-traffic areas while extending service to underserved regions.

The Emergence of For-Hire Vehicles (FHVs) and Ride-Hailing Platforms

The for-hire vehicle (FHV) sector has long complemented NYC's taxi system, traditionally serving pre-arranged trips through livery cars and black cars. These vehicles cater to passengers seeking more personalized service than standard street-hail taxis, particularly in outer boroughs and less-trafficked areas. However, the early 2010s ushered in a transformative era with the rise of app-based ride-hailing platforms like Uber and Lyft, fundamentally altering urban mobility dynamics. Unlike traditional taxis, FHVs operate using app-based algorithms to calculate fares dynamically based on factors such as real-time demand, traffic conditions, and driver availability. This shift introduced innovative features like upfront fare estimation, GPS-based route optimization, and cashless payments, providing a seamless experience for passengers. Additionally, ride-hailing platforms diversified service offerings, including premium options like Uber Black and shared rides, expanding accessibility across different income levels and travel preferences.

- **4. Methodology and Findings**

This thesis employs a multidisciplinary methodology combining descriptive analysis, modeling techniques, and comparative methods to explore urban mobility patterns and fare elasticity across various transportation modes. The methods were selected to address different facets of the research, ranging from analyzing spatial and temporal trends to quantifying the impacts of operational and contextual factors on urban transportation systems.

Descriptive Analysis

Descriptive methods served as the foundation for analyzing and visualizing trends across datasets, offering insights into spatial, temporal, and demographic patterns relevant to transportation systems. These methods were applied universally across all research components to provide a coherent understanding of the underlying data.

Temporal analyses involved exploring trends in trip volume, duration, and distance across different times of the day and week. Line graphs and histograms highlighted variations in user behavior, peak usage times, and operational dynamics. Spatial patterns were analyzed using geospatial mapping techniques, which visualized the distribution of trips across pickup and drop-off zones, emphasizing regional differences in service utilization.

While these methods were broadly applied, some descriptive techniques were tailored to specific analyses. For example, demographic and geographic factors were integrated into the analysis using supplementary datasets, such as census data and public transportation accessibility measures. Population density and transit accessibility metrics were spatially matched to transportation data, providing additional context for understanding demand and supply distributions across urban areas.

Modeling Techniques

Statistical and computational models were employed to quantify relationships between variables and assess transportation dynamics. These models offered insights into factors driving demand, pricing, and mobility patterns across multiple modes of urban transit.

Ordinary Least Squares (OLS) Regression

OLS regression was a common tool applied across multiple research components to identify the factors influencing trip characteristics and fares. Common independent variables included trip distance, duration, traffic conditions, temporal indicators (e.g., weekends and rush hours), and location-based attributes. These models established baseline relationships, allowing for comparisons between modes such as buses, ride-hailing services, and traditional taxis.

Fixed Effects Models

To account for unobserved heterogeneity in spatial and operational characteristics, fixed-effects (FE) models were utilized. These models incorporated fixed effects for locations, such as pickup or drop-off zones, to control for location-specific influences while isolating the impact of demand factors like congestion and time of day. FE models were particularly useful in identifying how urban density and geographic variations shaped pricing and demand responsiveness across transportation systems.

Cluster-Based Analyses

K-means clustering was used to segment locations into categories based on shared characteristics, such as urban density or transit accessibility. This approach provided a more granular understanding of transportation patterns by grouping zones into clusters, such as high-density

pickup to low-density drop-off areas, allowing for detailed insights into fare and demand variations across different urban contexts.

Advanced Modeling Approaches

In parts of the thesis requiring predictions or analysis of sequential data, machine learning models were applied. Long Short-Term Memory (LSTM) networks captured temporal dependencies in hourly and daily demand trends, while Gradient Boosting techniques uncovered non-linear relationships among operational and contextual factors. These advanced models complemented traditional methods, offering a more comprehensive understanding of transportation dynamics.

Comparative Methods

Comparative analyses were integral to understanding differences between transportation modes and their operational frameworks. Descriptive statistics compared attributes such as trip distances, durations, and fares across modes, providing insights into user behavior and service efficiency. Statistical significance tests, including t-tests and Mann-Whitney U tests, were applied to identify meaningful differences in observed trends.

Geospatial comparisons further explored regional disparities in service usage. Maps illustrating the dominance of specific transportation modes in different zones revealed spatial variations in accessibility and demand. For example, some analyses focused on relative concentrations of public transit versus taxis to identify gaps in service coverage or to assess the complementarity of modes. The methodology was designed to address a range of research questions across individual parts while maintaining coherence in the overall analysis. Descriptive methods provided a foundation for exploring trends and patterns, while regression and clustering techniques quantified relationships and identified nuanced differences between modes. Advanced models enhanced the

depth of analysis, offering insights into temporal dynamics and complex interactions. Together, these approaches provided a holistic view of urban mobility, capturing both the shared and distinct characteristics of fixed and dynamic pricing systems.

- **5. Introduction to Individual Parts**

In this research, a mix of predictive modeling, geospatial analysis, and qualitative methods have been employed: LSTM networks and Gradient Boosting to uncover temporal patterns, fare trends, and feature interactions. Geospatial analysis mapped demand density, accessibility, and transit gaps across boroughs. Comparative studies evaluated operational and pricing differences between taxis, ride-hailing services, and public transit. Public transit impacts were analyzed through correlations with shifts in taxi and ride-hailing usage. Additionally, socioeconomic factors and policies were integrated to contextualize findings, such as fare caps and congestion pricing, along with weather conditions and traffic, which further enriched the analysis, providing actionable insights on cost efficiency and accessibility.

The analysis begins by examining mobility trends in NYC's outer boroughs following the introduction of enhanced taxi coverage, such as the Green Taxi program. This research addresses the question of whether these measures have effectively improved transportation equity by increasing access to reliable taxi services in historically underserved areas and how these changes have influenced broader mobility patterns.

Next, the relationship between NYC's public bus system and taxi usage is explored, with a focus on understanding the interplay between these modes of transportation. By analyzing fluctuations in taxi demand relative to bus service enhancements and operational changes, this section seeks to

uncover whether improved public transit services can alleviate taxi dependency or whether they coexist in a complementary fashion.

Building on this, the variability in costs between comparable Green and Yellow taxi rides is studied over time, addressing questions about fare consistency and the factors driving price differences between these two services. This analysis showcased how difficult it is, for comparable matching rides, to study the variability of fares over time, and how, in order to obtain a fare cost difference prediction, it is better to rely not just on geospatial data but on other complementary elements as well.

The next part investigates trip purposes and usage patterns, comparing Green Taxi users with ride-hailing service users such as Uber and Lyft. This section explores the distinct user behaviors and preferences that shape demand for these services, answering questions about how spatial, temporal, and demographic factors influence trip purposes and usage frequency.

Finally, the analysis turns to pricing models, focusing on the comparison between fixed pricing (used by traditional taxis) and dynamic pricing (employed by ride-hailing platforms). By investigating fare sensitivity to demand-driven factors such as traffic and weekends, this section addresses how pricing mechanisms impact urban mobility and service accessibility.

- **8. By Massimo Pio Carollo: A study on variability of cost difference between Green and Yellow NYC taxi rides over time.**

8.1 Introduction

Since the early 20th century taxis have been a vital mode of transportation for New Yorkers. The first yellow taxis introduced in the 1920s became quickly a symbol of the city, offering metered rides to passengers across the five boroughs. Over the decades the taxi industry expanded and adapted to changing urban dynamics; until in 2013 when the city introduced green taxis, known also as Boro Taxis: the decision was taken to address the lack of taxi in underserved areas. The difference between them relies in the operating areas: Yellow works in in all five boroughs of New York City, (Manhattan, the Bronx, Queens, Brooklyn, and Staten Island) and central airports, Green taxis serve the same areas, but are prohibited to respond to street hails in the city's core high-demand taxi zones (below 96th Street on the East Side and 110th Street on the West Side) where yellow taxis have their dominance.

So, except from the core of Manhattan, both Green and Yellow taxis provides a comparable mobility service.

- How the service provided differs in terms of costs between yellow and green taxis rides?
- How does this difference variates over time?
- Which prediction model (Long Short Term Memory and Gradient Boosting) suits the best data to better understand and predict the cost difference values?

In order to obtain valid conclusions, we calculated the costs difference between green and yellow taxi rides up to a specific match criteria: the 2 rides in order to be compared have to share the same time and location pick up and drop off coordinates: both of them must happen in a gap of max one

hour one from each other and have to start and finish at the same location (in this study the same taxi zone; those areas correspond to different municipalities on new York city). Additionally, in order to study temporal variability, the cost difference values have been aggregated in daily and hourly average values.

The results of this investigation shown that the variability of cost differences between rides doesn't follow neither strong specific temporal and location patterns, even with adding engineered temporal features: We searched for hourly and daily patterns performing Long Short-Term Memory models. However, the predictions results, put in the context of variance and deviation of values found lead to the conclusion that the model struggles with denser parts of the data and is not able to deal with the high variability of values.

This is why the Long-Short Term Memory model has been outperformed by a Gradient Boost prediction model, that predicted cost difference values over statistically significant variables inside the datasets, as the duration, distance and the tips of each ride; achieving a significantly lower Mean Squared Error and provided a way more accurate predictions and insights into the evolution of cost differences over time.

8.2 Literature Review

The cost of taxi rides has changed a lot over the years, due to changes in policy, market dynamics and the entrance into the market itself of new players like Uber and Lyft. Consequently, during the previous decade, we observed a significant decline in the number of taxi rides and an increase of competition between taxis and app-based third-party services (Qian & Ukkusuri, 2017), since competitor's services and availability, beside a high fare, outperform taxis in service quality (Sun et al., 2023).

The new entry “ride-hailing” services played a key role in fare increases: Taxi rates are now higher in real terms than before deregulation, that often lead to monopolistic or oligopolistic pricing behaviors rather than competitive one’s deregulation (Teal & Berglund, 1987).

There is a wide literature about studies on cost evolutions on taxis, and about all the factors that can influence the variability on the final cost of a ride. Taxi fares depend on various factors including drop charges, duration charges and linearly increase with trip length due to traffic. (Yang, Ye, Tang, & Wong, 2005), with a consequent impact on both demand and pricing strategies of ride-sourcing platforms.

The idea of a study with at the base a fare comparison with origin-destination and same time window is not new: Uber and taxi fares has been compared covering 277,840 rides over 10 months analyzing fares based on trip characteristics as distance, duration, and time for 40 origin-destination (Rangel et al., 2021). Taxi fares were analyzed using official city regulations, enabling a consistent comparison under identical trip conditions (Yang et al. 2021). Between comparable rides, App-Based Ride Services were found to be 40% cheaper than taxis for the same origin-destination pairs. Additionally, ride-hailing services had significantly lower wait timestamp a near-guaranteed pickup rate versus taxis, where 19.5% of riders were not picked up. (Brown and LaValle 2020). Globally, the Association of Public Transport (UITP) compares every year taxi rides performances of 16 cities worldwide, including rides fares, showing variability between cities due to local regulatory frameworks and market dynamics (UITP, 2020).

Regarding trend analysis, there are present studies that calculate average daily passenger demand and the consequences on its increase or decrease, as the passenger demand and the number of licensed taxis. About fares, excessive hikes significantly reduce demand and improve service quality by increasing taxi availability (Hai Yang et al. 2000). Additionally, an adjusted fare of taxi

rides can lead to higher profits under specific conditions of market demand elasticity (Li & Szeto, 2021).

Other studies about dynamic pricing comparing Uber and Taxi rides suggest that the variability is influenced by trip distance, duration, delays, and weather conditions and shows a correlation between fares rises fares and peak hours along with high-demand periods like weekends and holidays (Rangel et al., 2021). Additionally, is shown that Uber's frequent tariff changes contrast with taxis that adjust fares less often. As for New York, Yellow Taxi fares remained unchanged for eight years before an increase in 2012(Noulas et al., 2017).

Long-short term memory (LSTM) models have already been used in the past dealing with taxi data to predicting yellow taxi demand in New York City: firstly, a linear regression was performed to understand explanatory variables like day of the week, holidays and weather then, the LSTM is applied to handle non-linear trends (Kim et al., 2020). Since the goal of this study is to analyze temporal patterns, LSTM model appear to be the right tool to do so: it is one of the most advanced networks to process temporal sequences, as it is specially designed to overcome the exploding/vanishing gradient problems that typically arise when learning long-term dependencies (Van Houdt et al., 2020). To help the model to perform and predict accurately is possible to create and use engineered features: they can help summarize the dynamics of the data and enrich its representation (Cerqueira et al., 2024).

In this research we are going to understand the evolution of difference in terms of cost of two similar services (yellow and green taxis), using the Long Short Term Memory approach, and to understand possible relation with other features.

Based on previous work with the same method applied (Hassan et al., 2023), the LSTM can effectively model complex dependencies between features like trip distance, time of day, and

weather conditions: the use of spatiotemporal external factors will enhance fare prediction accuracy. Additionally, LSTM's proven performance in reducing rolling errors indicates robust long-term predictive capabilities without frequent retraining makes it the right choice for dynamic fare trends.

To study the general evolution of the cost difference, beside the time relations, Gradient Boosting Models are a powerful machine learning technique commonly used combined with different machine learning approaches to address specific challenges in taxi demand prediction (Vanichrujee et al., 2021). Gradient Boosting models are commonly used in the demand and fair prediction of taxi thanks to the ability to handle multivariate inputs and account for the spatial-temporal variability in demand, (Poongodi et al., 2022). usually consistently outperformed other models in terms of precision, recall, and F1 score (Rajendran et al., 2021).

8.3 Methodology

Data preprocessing

In order to reach the goal of understanding cost variability and having trustable costs models, identifying key variables is a crucial step: this has been done by preprocess different datasets and handling only the relevant information from them. The datasets available on the Taxi and Limousine Commission of New York City website (https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page) provide, divided for each month, yellow and green taxis rides distinct datasets in parquet format. The datasets comprehend for each ride data about the pick-up/drop-off locations and datetime, along with multiple cost voices, tips, right leght and other useful variables. Considering the size of each file, (more than 3000000 rows for each of the 36 files: 18 from yellow and other 18 for green taxi rides monthly files from January 2023 to June 2024), all the files have

been manipulated thru iterative loops, to a faster and more precise control of the operation of processing data.

Data Cleaning

Initial datasets had issues such as missing values, zero distances or costs, and presence of outliers. The data cleaning pipeline addressed these issues systematically.

Missing values were removed, the dataset filtered by date ensured data consistency. Invalid trip durations, such as zero or excessively long trips, were excluded. Trips with zero distances or costs were removed. Outliers were removed as well using statistical thresholds applied to trip distance, duration, total cost, and tips.

Database features (before matching green and yellow rides)

To later better study the correlation between all the different features regarding every single taxi ride, and consequently to obtain later different predictional results, we added additional columns to the datasets: $TripDuration_i = \frac{DropoffDateTime_i - PickupDateTime_i}{3600}$

TripDuration is calculated by the difference between DropoffDateTime and PickupDateTime in seconds converted to hours by dividing by 3600, resulting in unit/hours. The range of values of all values is 6 minutes to 3 hours.

$$TotalCost = \sum_{i=1}^{n} (FareAmount_i + Extra_i + MTATax_i + TollsAmount_i + AirportFee_i)$$

TotalCost is measured in dollars and represent the sum of FareAmount, Extra, MTATax, TollsAmount, and AirportFee. The range boundaries of TotalCost are from 1 to 58$.

Matching and additional features

After data cleaning and preparation of both green and yellow datasets, we merged each yellow monthly taxi rides file with the correspondent green taxi monthly rides file. 2 rides to be matched, and then used for subsequent analysis must share the same pickup and drop off location, the same month, day and both of them must happen in a gap of max one hour one from each other.

The criteria used to merge and match rides allow us to ensure fair comparison. This controls minimize temporal/spatial variability and the influence of external factors, therefore will be possible to focus on key operational metrics such as trip costs, durations, distances, and tips for both green and yellow taxis.

Moreover, we have created additional variability features generated from already available data:

DistanceDifference = TripDistanceYellow – TripDistanceGreen

DurationDifference = TripDurationYellow – TripDurationGreen

CostDifference = TotalCostYellow – TotalCostGreen

TipsDifference = TipsYellow - TipsGreen

These variables represent the differences between operational metrics of yellow and green taxis variations regarding trip distance, duration, total cost, and tips. They quantify the disparities in service performance and economic outcomes between the two taxi types. Those features will be the base for future analysis and to predict the factors driving these differences. Our target variable will be CostDifference: it represents for each match the difference in dollars between the total fair value of the yellow ride and the total of the green one.

Cyclical variables

In order to represent cyclical transformations of temporal data we added: pickup_hour_sin, pickup_hour_cos, pickup_weekday_sin, and pickup_weekday_cos: specifically, from the hour of

the day and the day of the week for the pickup time. These variables capture the circularl nature of time and are crucial for modeling and predicting cyclical patterns.

To address this, sine and cosine transformations were applied.

$$sin\_value = sin\left(\frac{2\pi \cdot time\_unit}{max\_time\_unit}\right), cos\_value = cos\left(\frac{2\pi \cdot time\_unit}{max\_time\_unit}\right)$$

here max timeunit is 24 for pick-up hour and max timeunit is 7 for pick_up weekdays.

In the table below is possible to see the correlational values of the final dataset that we are going to use for modeling.

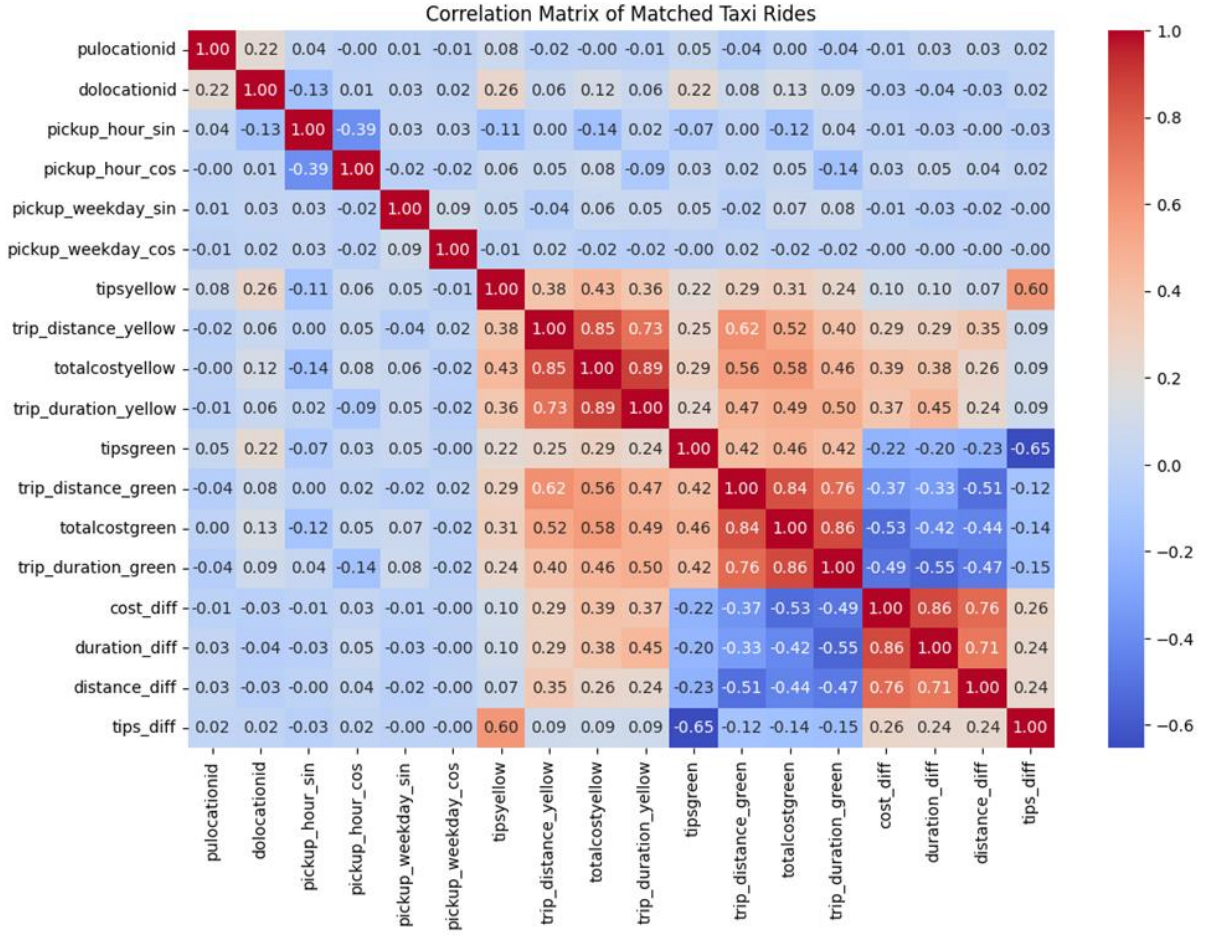

Figure 1. The correlational matrix of the matching rides dataset

8.4 Models: Long Short Term Memory and Gradient boost model

To better understand hourly and daily trends, we aggregated and created two subset of matching rides dataset, respectively the first for hourly and the second for daily average values, allowing us to perform 2 different analysis using different timestamps: we obtained a daily dataset with 546 rows and another with 13104 rows (total number of days and hours in the time window analysed). In order to study daily and hourly average trends of cost difference we used Long Short-Term Memory (LSTM) recurrent neural network: it is designed to handle sequential data and learn dependencies over time, thanks to its architecture the model overcome struggles with retain long-term dependencies due to vanishing or exploding gradient problems.

Long Short Term Memory model uses a memory cell structure to select if to retain or forget information. The core mechanism involves three gates that regulate the flow of data: input, forget, and output:

The forget gate f_t determines whether historical trends in the target variable cost_difference are relevant for predict its future ones, with a specific (daily or hourly) timestamp. It is defined as $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ where σ is the sigmoid activation function, W_f represents the weights matrices, $[h_{t-1}, x_t]$ is the previous hidden state from the previous timestamp, x_t is the current input, in our cases daily/hourly averages of pickup_hour_sin, pickup_hour_cos and the other variables b_f, is the bias term.

Then the input gate decides which new information to add to the memory is formed by two components: The Input gate activation $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ and the candidate memory $\widetilde{C_t}$ that calculates potential new memory content defined as $\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$ with weights matrices for the input gate and candidate memory. Together they control how new information updates the memory cell.

The output gate determines the part of the memory to influence the current output: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$where the hidden state is computed as $h_t = o_t \cdot \tanh(C_t)$ with h_t representing relevant time-dependent information from the past and current days/hours for predicting cost_difference. The memory C_t cell state is updated as: $C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t}$ where$f_t \cdot C_{t-1}$ is the retained memory that retains relevant information from past timestamps like weekly patterns for cost_difference and $i_t \cdot \widetilde{C_t}$ is the new memory added with new information such as spikes in the data.

The hidden state h_t is passed to a dense fully connected layer to generate the predicted value of cost_difference: $cost\_diff = \text{Dense}(h_t)$. During training, the model compares the predicted values of cost_difference $y_{\text{pred},i}$ with the true ones $y_{\text{true},i}$ to compute the error: $\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{\text{true},i} - y_{\text{pred},i}\right)^2$.

Lastly Gradients of the loss are backpropagated through the LSTM's gates and cell states to update the weights, improving the model's ability to predict cost_difference over time.

Both the Long Short Term Memory models performed have a similar structure: two layers with 50 units each. 80% of the data is trained with the gradient-based optimization algorithm Adam. The difference relies in timestamps of 6 (hours) for the hourly and 7 (days) for the daily model and the butch size (50 for hourly and 10 for daily).

Long Short Term Memory Results

The Long Short Term Memory Hourly model achieved a Mean Squared Error score of 5.260: this value is moderate compared to the variance of 6.19 and standard deviation of 2.49, Therefore we can state that the model captures the variability of the target but still can be improved. The model struggles with denser parts of the data since the Mean Absolute Error (1.388) exceeds the

Interquartile Range (1.82). Errors are relatively small compared to the range of 25.45. Seeing all those statistics is possible to say that the model performs reasonably well, but the level of accuracy of performance could be improved.

The second Long Short Term Memory model performed with the aggregated Daily dataset. With a variance of 0.2669 and standard deviation of 0.5167, the Mean Square Error of 0.12524 and Mean Absolute Error of 0.2695 showcase how even in the densest data the model performs well. Even considering that the MAE is smaller than the Interquartile range of 0.4116. Considering the target's range (9.797), Errors are minimal and shows good accuracy. The negative Coefficient of variation -61.84% reflects a low or near-zero mean but does not detract from the model's strong predictive performance. The results overall suggest effective modeling with small errors accountable to data variability.

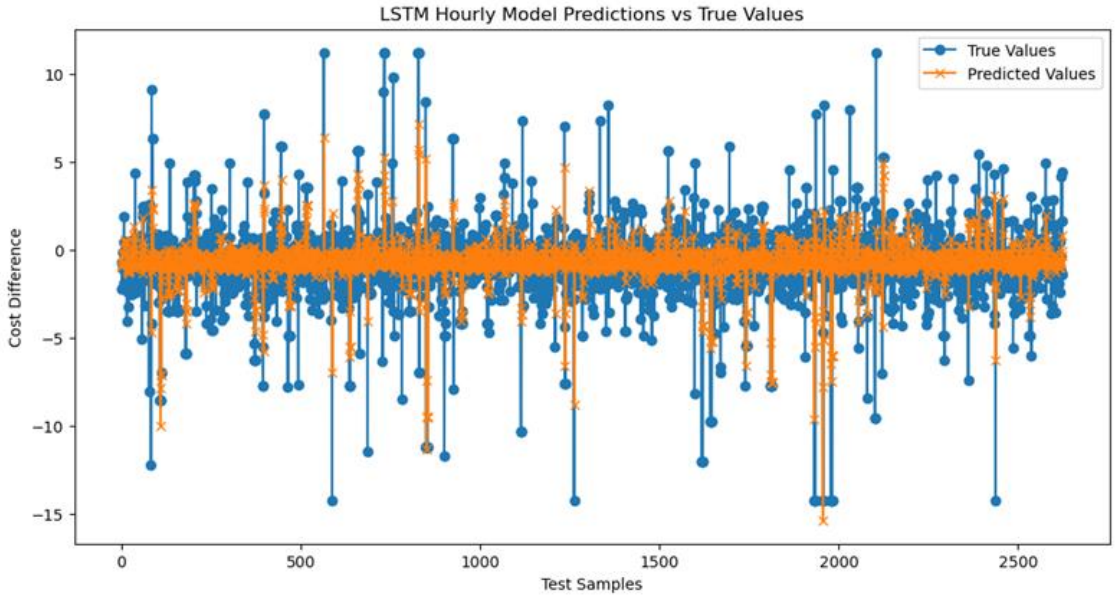

Figure 2. Long Short-Term Memory Hourly model predicted results for Cost difference

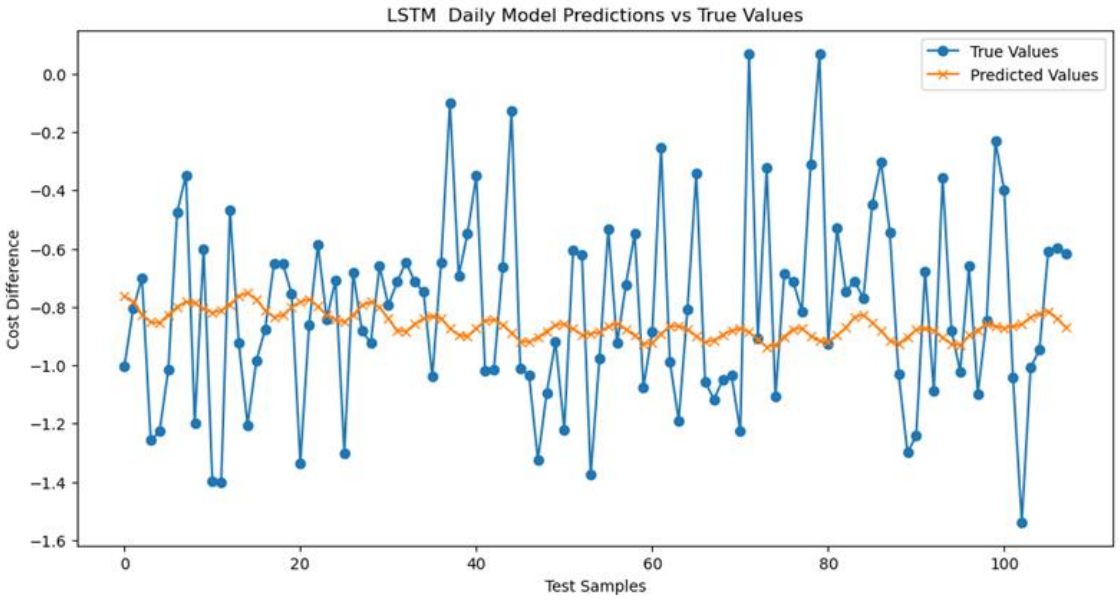

Figure 3. Long Short-Term Memory Daily model predicted results for Cost Difference

Given the results and the lack of strong temporal patterns for cost_difference, we analysed the matching files datasets without focusing on temporal patterns, trying to predict the values of cost_difference considering a different approach: The extreme Gradient Boosting model resulted as a more suitable choice: the Gradient Boost model focuses on a global analysis of features relationships and selects the most important ones instead of focusing on temporal dependencies. This allows to study non-linear interactions between variable that shown a higher correlation with cost_difference: (tips_diff, distance_diff, and duration_diff) to improve prediction performance on overall values. The model is also robust with minimizing the impact of outliers. This aligns with the data characteristics and addresses the limitations of Long Short Term Memory for this analysis.

Extreme Gradient Boost model

Extreme Gradient is a boosting algorithm that builds an ensemble of decision trees composed by a loss function and a regularization term:

$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$.

$l(y_i, \hat{y}_i)$ represent the loss function. It can be defined as $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ with y_i as the actual value of cost_difference and $\hat{y}_i$ the predicted one, based on tips_diff, distance_diff, and duration_diff.

$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda|w|^2$ allows to penalize tree complexity, whit T as the number of terminal nodes in the decision tree, λ is a regularization parameter that penalizes the magnitude of leaf weights, and γ to penalizes the number of leaves. The aim of those variables is to together ensure the model does not overfit.

In each iteration a new tree $f_t(x)$ is added to correct the residuals error $r_i^{(t-1)}$ in predicting cost_difference from the prior trees using tips_diff, distance_diff, and duration_diff: they are computed as the negative gradient of the loss function $r_i^{(t)} = -\frac{\partial l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}}$

The model updates predictions as $\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta f_t(x_i)$, where η controls how much each tree contributes. Leaf weights are optimized using the gradient g_i and the second-order derivative *h(i)* of the loss function $w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$ with I_j as set of data points assigned to leaf j , g_i is the gradient of the loss function for cost_difference with respect to predictions, h_i is the second-order derivative of the loss function (Hessian) and λ controls the regularization of the weights.

Gradient Boost determines whether splitting on tips_diff, distance_diff, or duration_diff improves the model by considering gradients and Hessians of cost_difference for the corresponding data points.

Once the model is trained, predictions for cost_difference are computed as: $y_i = \sum_{t=1}^{T} \eta f_t(x_i)$, with *T* as the total number of trees: each tree contributes a weighted prediction based on relationships learned between tips_diff, distance_diff, and duration_diff.

Gradient Boost Results

The Extreme Gradient Boost model trained on a dataset of 905,121 samples and 22 features; achieved a Mean Squared Error of 4.33, lower than the target's variance of 5.96 and reasonable compared to its respective range of 25.45. The model shows strong predictive power with an R^2 score value of 0.8067. Predictions are generally close to actual values Predicted values are generally close to the actual ones, but with occurring of some deviations. With an interquartile Range of 1.72 and Standard Deviation of 2.44, the target shows moderate variability. Seen so we can say that the model effectively captures the target's patterns, with room for further small optimizations.

Sample Predictions vs Actuals Cost Difference Values with the Gradient Boost Model:

Predicted: [4.300309], Actual: [4.9]

Predicted: [0.44483578], Actual: [-3.05]

Predicted: [-0.3078202], Actual: [0.]

Predicted: [8.367459], Actual: [8.8]

Predicted: [0.13561726], Actual: [-1.]

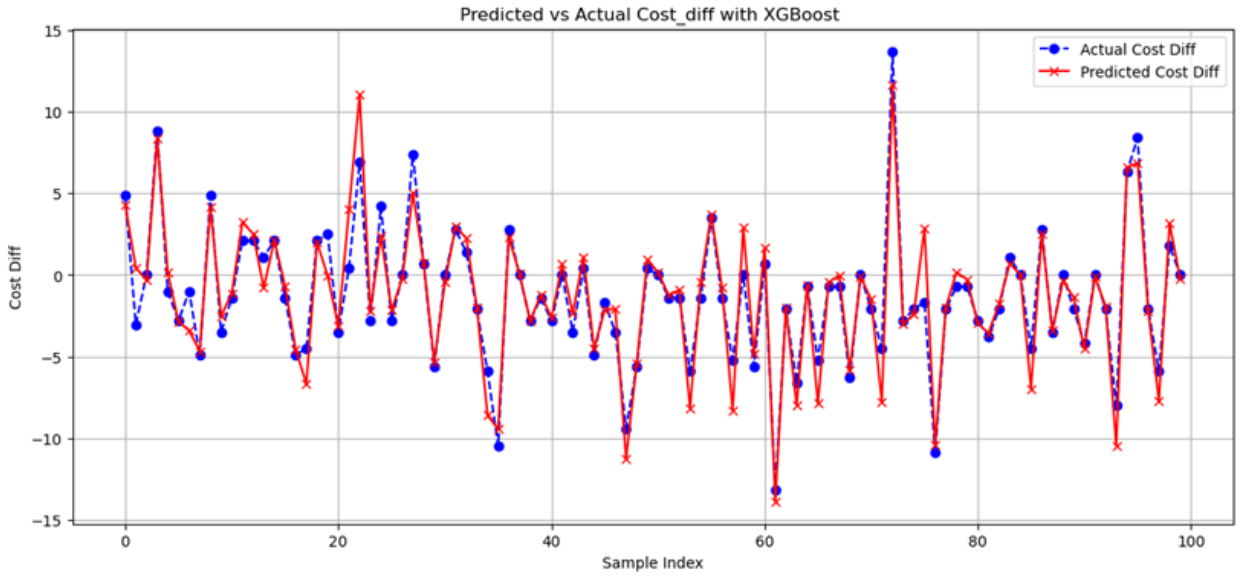

Figure 4. Visualization of Gradient Boosting prediction model of Cost_difference

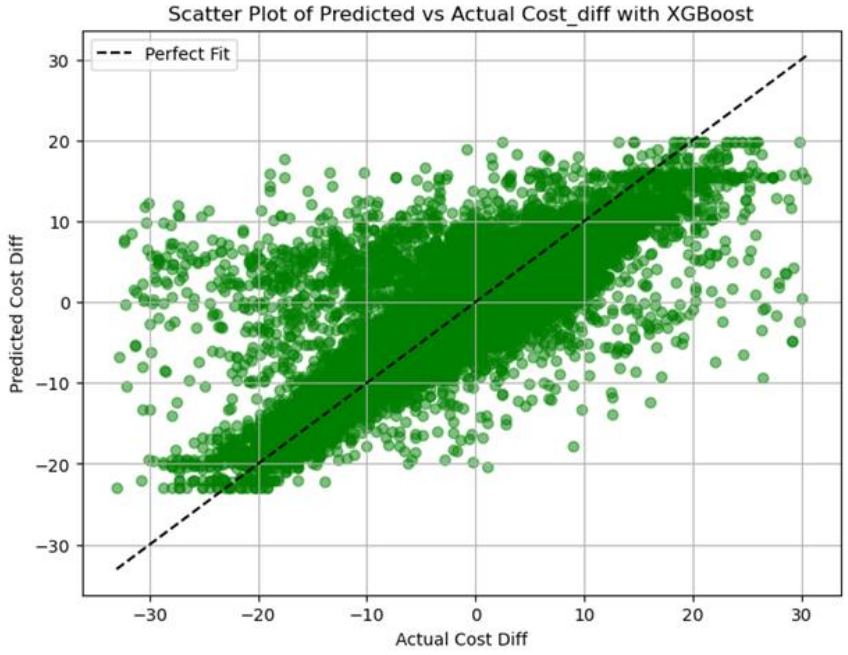

Figure 5. Scatter plot of predicted values versus actual values for a model predicting Cost Difference (Cost_difference) using Gradient Boosting

8.5 Conclusions

This research has been able to underline the importance of aligning modeling techniques with data characteristics. This study investigated the variability of cost differences between Yellow and

Green NYC taxi rides. with Long Short-Term Memory (LSTM) models with the goal of capturing hourly and daily trends: It has demonstrated limited success in detecting temporal patterns due to the inherent high variability and weak correlations with time-based features. Even if the daily model achieved satisfactory results for general trend analysis, the hourly model struggled with dense data regions underlying the dataset's complexity.

To address these limitations another model with Extreme Gradient Boosting was performed without focusing anymore into aggregating data following temporal patterns, but on feature interactions and dependencies.

The gradient Boosting model achieved significantly lower error metrics and therefore demonstrated superior predictive accuracy and robustness: This approach was able to capture effectively the relationships between the variability of cost_difference and key variables as trip distance, duration, and tips.

Future work could explore additional features of engineering along with including additional data for deeper and more complete understanding of other factors able to influence the cost variability. Additionally, others prediction models can be performed to further understand and improve predictive performances.

13. Resources

- Brown, Annxe, and Whitney LaValle. "Hailing a Change: Comparing Taxi and Ridehail Service Quality in Los Angeles." *Transportation* 48, no. 2 (February 10, 2020): 1007–31. https://doi.org/10.1007/s11116-020-10086-z.
- Centre for Cities. 2016. “Transport Essential for Growth in Cities - Centre for Cities.” October 14, 2016. https://www.centreforcities.org/reader/delivering-change-making-transport-work-for-cities/transport-essential-growth-cities.
- Cerqueira, Vitor, Nuno Moniz, and Carlos Soares. "VEST: Automatic Feature Engineering for Forecasting." *Machine Learning* 113, no. 7 (April 6, 2021): 4523–45. https://doi.org/10.1007/s10994-021-05959-y.
- Cervero, R. "The Transit Metropolis: A Global Inquiry into the Transportation–Land Use Nexus." *Island Press* (1998). https://search.worldcat.org/title/502343506?oclcNum=502343506.
- Chang, Hung-Hao. 2017. “The Economic Effects of Uber on Taxi Drivers in Taiwan.” *Journal of Competition Law & Economics* 13 (3): 475–500. https://doi.org/10.1093/joclec/nhx017.
- Contreras, Seth D., and Alexander Paz. 2017. “The Effects of Ride-Hailing Companies on the Taxicab Industry in Las Vegas, Nevada.” *Transportation Research Part A: Policy and Practice* 115 (November): 63–70. https://doi.org/10.1016/j.tra.2017.11.008.
- “Employment Patterns in New York City Trends in a Growing Economy.” 2016. Uploaded by NYC Planning Commission. Nyc.Gov. New York City Planning.

https://www.nyc.gov/assets/planning/download/pdf/data-maps/nyc-economy/employment-patterns-nyc.pdf.

- "The Impacts of Taxicab Deregulation in the USA on JSTOR." *Www.Jstor.Org*. https://www.jstor.org/stable/20052801.
- "$100 Billion Cost of Traffic Congestion in Metro New York - Partnership for New York City." 2020. Partnership for New York City. February 20, 2020. https://pfnyc.org/research/100-billion-cost-of-traffic-congestion-in-metro-new-york/