Business Survival Rate in NYC:

Focusing on Median household Income, Population and Accessibility

Qingyuan Feng, Mingchuan Ma, Xianze Gao, Jiuling Zhong

New York University

Center for Urban Science + Progress

CUSP 7033 Machine Learning for Cities

Abstract

This research delves into what makes businesses successful, in New York City looking at how median household income, population density and accessibility play a role. By using a range of data and sophisticated analysis methods the study reveals that areas with households tend to see more commercial activity. It also shows that populated areas offer both customers and tougher competition for businesses. The ease of reaching transportation centers and the availability of parking spaces are factors influencing the success of ventures. The findings from this study offer insights for policymakers and city planners, on fostering growth in urban settings.

Keywords：Urban Economics, Median Household Income, Business Viability, Spatial Analysis, Machine Learning

Introduction

Urban environments are complex systems where economic dynamics play a critical role in determining the viability and success of businesses. In New York City, a metropolis renowned for its diverse economic landscape and high population density, understanding these dynamics becomes even more crucial. The city's heterogeneous economic conditions across its boroughs present a unique opportunity to examine how various factors, including median household income, population density, and accessibility, influence business survival.

This study aims to delve into the urban economic fabric of New York City by employing advanced machine learning techniques to analyze a comprehensive dataset. Models such as Random Forest, Support Vector Machine (SVM), and Naive Bayes are utilized to interpret a broad range of data inputs from multiple city databases. This approach enables a nuanced understanding of the interplay between different urban factors and business viability.

Recent data from the Census Bureau (2022) highlight the significant impact of median household income on economic activities, where a decrease in median household income has been linked to broader economic challenges, including increased poverty rates. This relationship underscores the importance of economic stability as a foundation for business success. Furthermore, the distribution of income and economic trends from various reports suggest that areas with higher income levels tend to support more robust business environments.

Population density is another critical factor. High-density areas offer a substantial customer base, which is vital for business sustainability. However, these areas also present challenges, such as increased competition and higher costs of operation. Businesses must navigate these complexities to succeed. The strategic placement of businesses near transit hubs can mitigate some accessibility issues, enhancing customer inflow and business performance.

Accessibility, particularly in terms of proximity to transport hubs and availability of parking, significantly impacts business success. Areas well-served by public transportation, such as subway lines and bus stops, provide easier access for both customers and employees, which can enhance business performance. In contrast, adequate parking facilities are crucial in areas where public transportation is less prevalent, as they significantly influence customer choices and business viability.

By integrating these insights with machine learning analysis, this research aims to provide stakeholders and urban planners with predictive tools and strategic insights to foster sustainable urban economic development. The application of such technologies in urban studies is pivotal, not just for academic inquiry but for practical policy-making and urban management to adapt to the dynamic changes in city environments.

These enhanced references provide a solid foundation for understanding the interdependencies of economic factors and their impacts on urban business viability. In summary, this study seeks to enhance our understanding of the interdependencies among economic factors, population dynamics, and accessibility in urban settings, offering valuable guidance for creating a conducive environment for business success in metropolitan areas like New York City.

Literature Review

In recent years, the correlation between economic factors and the survival rate of businesses has been extensively researched, particularly in urban settings. As cities continue to expand, understanding the dynamics that influence business survival and growth becomes increasingly critical. New York City, with its diverse economic landscape and high population density, serves as an ideal case study for examining these relationships.

Studies consistently highlight the role of median household income and population density as pivotal determinants of business success. Higher median incomes often indicate greater consumer spending power, positively impacting local businesses by increasing sales and profitability (Furman Center for Real Estate and Urban Policy, n.d.). This relationship is supported by data showing that areas with higher median incomes typically experience more robust business activities (Furman Center for Real Estate and Urban Policy, n.d.).

Similarly, population density plays a crucial role. Areas with higher population densities provide a larger customer base, vital for business sustainability (United States Census Bureau, 2020). However, these areas might also entail more competition and higher operational costs, leading to a complex balance that businesses must navigate (United States Census Bureau, 2020).

Accessibility, particularly in terms of proximity to transport hubs and parking availability, is another crucial factor affecting business performance. The presence of subway lines and bus stops enhances customer and employee access, potentially boosting business performance significantly (Department of Consumer and Worker Protection, 2024). Conversely, in areas where public transport is less utilized, adequate parking facilities become critical as they significantly influence customer choices and, consequently, business viability (NYC Planimetric Database, n.d.).

Moreover, the integration of advanced technologies, such as machine learning models like SVM, Naive Bayes, and Random Forest, in predicting business outcomes based on these variables underscores the complexity of urban economic environments. Studies by Pathak et al. (2023) and Sarker (2021) reveal the strengths and limitations of these models in handling overlapping data labels and processing large datasets. These insights guide the selection of appropriate analytical tools for urban economic studies, enabling researchers and policymakers to make more informed decisions.

Methods.

Median Household Income.

Four primary datasets are utilized in this project, each playing a crucial role in our analysis of business viability in New York City. The first dataset, Median Household Income, is derived from the 2020 Census and provides detailed records of median household income across various neighborhoods and districts within New York City based on census geo-locations. This dataset is essential as it highlights the economic diversity across different areas, enabling us to correlate income levels with business activity and survival rates. However, a significant limitation of this dataset is that it contains only the GEOID (Geographic Identifier) and lacks geometric information, such as the actual shapes and boundaries of the census tracts. To overcome this limitation, we integrated the median household income data with additional geospatial data from the census, which includes the necessary geometric information. This integration allowed us to create detailed visualizations on a map, illustrating how income levels vary across the city and providing a clearer picture of the economic landscape. These visualizations are instrumental in identifying patterns and trends in business success relative to income distribution, offering valuable insights for stakeholders and policymakers aiming to foster economic growth and business development in New York City

Business License Activity

The second dataset tracks business license activity and is crucial for determining the operational status of companies within New York City, including whether they are still active or have declared bankruptcy. This dataset comprises two main components: one detailing legally operating businesses with active licenses, and another containing data on license applications. By combining these two components, we gain a comprehensive view of business activities in the city. The active licenses dataset provides information about businesses that are currently authorized to operate, while the license application dataset offers insights into new businesses entering the market and those that may have ceased operations. To merge these datasets effectively, we used license numbers as a common identifier, ensuring accurate linkage of information. The data were sourced from the NYC Open Data platform, which is a valuable repository of publicly accessible data. By integrating these datasets, we can analyze trends in business openings and closures, assess the health of various industries, and identify factors that contribute to business longevity and success in different parts of the city. This comprehensive approach allows us to paint a detailed picture of the business landscape in New York City, providing stakeholders with critical information for making informed decisions about economic development and business support initiatives.

Accessibility

The third dataset focuses on accessibility, a crucial factor in determining business success in urban environments. In this study, accessibility is defined by three key parameters: the number of parking spaces, subway stations, and bus stations. Each of these elements plays a significant role in making businesses easily reachable for customers and employees alike. All three sets of data were sourced from the NYC Open Data platform, ensuring comprehensive and up-to-date information.

To analyze this data effectively, we combined the different accessibility parameters using spatial join functions, aligning them with census tract geo-locations. This method allows us to accurately map out accessibility features within each census tract, providing a detailed view of how easily accessible different areas are. Although parking lot areas are represented as polygons, their relatively small size compared to the larger census geospatial units enables us to integrate them seamlessly into the dataset. By mapping these accessibility factors, we can identify areas with high and low accessibility, offering insights into how these variables influence business viability and customer foot traffic. This integrated approach helps us understand the relationship between accessibility and business success, guiding urban planners and policymakers in making data-driven decisions to improve transportation infrastructure and support local businesses.

Population

Population data, also sourced from the 2020 Census, is based on census geo-locations and is well-organized. However, it contains several columns that were not relevant to our project, such as population by different age groups. To streamline the analysis, these extraneous columns were removed from the dataset.

After extensive cleaning and modification, we compiled a data frame where each row includes comprehensive details about a company: its license status, GEOID, parking area, subway station numbers, and bus station numbers.

Graphical Methods of Maps

In the second phase of our project, we prepared two visualizations. The first map visualizes the relationship between population distribution and industry distribution in New York City. We utilized a population distribution heatmap as the base map and overlaid it with markers representing different types of companies operating in NYC. Creating this map provided us with a general understanding of how population distribution might influence business activity in the city. We observed that most businesses are concentrated in areas with larger populations; however, some locations in Queens and Staten Island exhibit less business coverage.

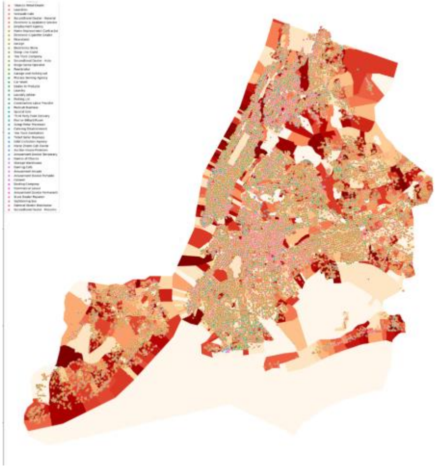

Figure 1. Population by Census Tract vs. Industry

The second map visualizes the relationship between median household income and industry distribution across New York City. Similar to the first map, this visualization offers a general perspective on how median household income may influence business activity. Our analysis revealed that areas with lower median household incomes have reduced business activities. This finding supports our hypothesis that lower median household incomes, indicating reduced purchasing power, lead to diminished business activity.

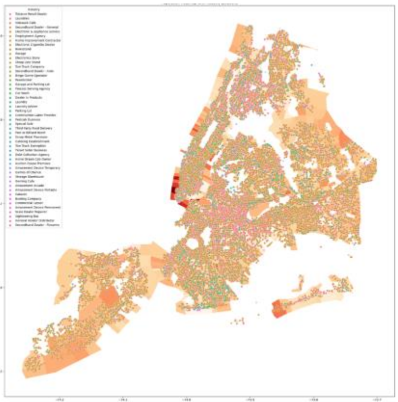

Figure 2. Median Income by Census Tract vs. Industry

Machine learning methods and techniques

In the third phase, we trained Random Forest, Support Vector Machine (SVM), and Naïve Bayes to unravel the connections that shape urban commercial environments. The datasets cover a range of factors from information to accessibility measures like proximity, to public transportation hubs and parking availability.

Random Forest, SVM, and Naive Bayes are robust algorithms that can effectively predict business survival rates based on factors like population, accessibility, and median household income. Random Forest handles non-linear data well and provides importance scores for each predictor, helping in understanding feature influence. Random Forest also can manage large data sets with mixed data types and effectively ranks the importance of variables, providing insights into which factors most impact business longevity. SVM is excellent for capturing complex relationships in data through its kernel trick, making it suitable for scenarios where the relationship between survival rate and inputs is not straightforward. Naive Bayes, meanwhile, offers fast predictions and performs well with categorical inputs, making it a good choice for preliminary analysis and when computational resources are limited. Together, these methods provide a comprehensive approach to modeling in diverse scenarios, enhancing the reliability of predictions. These three models could cover all the factors we are working on and give us a prediction on which factors will affect the most on the business survival rate.

Our model was developed to empower stakeholders and startup owners with actionable insights for establishing businesses in New York City. By evaluating the demographic characteristics across different geographical locations and utilizing the Random Forest algorithm, we identified key factors influencing business survival. Our findings highlighted that population density significantly affects business success, and the presence of competing businesses in close proximity also emerged as a critical factor. Interestingly, among accessibility metrics, the availability of nearby parking facilities was the sole factor with a significant impact on business survival rates. These insights are invaluable for prospective business owners, underscoring the critical importance of location demographics and accessibility within the urban landscape of NYC. By understanding these key factors, entrepreneurs can make informed decisions about where to establish their businesses, potentially enhancing their chances of long-term success. The model also provides valuable data for city planners and policymakers, helping them to design and implement strategies that support business growth and sustainability. Looking ahead, we plan to expand our model's predictive capabilities by integrating real-time data and machine learning advancements, while also analyzing the industry type of each business. We aim to incorporate additional variables such as economic trends, consumer behavior patterns, and the impact of online retail on brick-and-mortar establishments, enabling a more comprehensive analysis and providing even more nuanced insights. By continuously refining and updating the model, we hope to offer a dynamic tool that evolves with the changing urban environment, providing stakeholders with the most relevant and timely information to support their business endeavors.

Results.

During the evaluation of three distinct predictive models, diverse outcomes were noted. The SVM model, which included parameters such as population density and the density of similar nearby businesses—akin to those used in 2-D logistic regression—demonstrated clustering of scatter points when plotted. This clustering adversely affected the model's predictive accuracy. Furthermore, the survival rates projected by the model followed a linear pattern, contradicting our predictions. As noted by Krithika Suresh and her colleagues in their article, “Survival prediction models: an introduction to discrete-time modeling,” published in BMC Medical Research Methodology, machine learning algorithms are invaluable for their ability to discern non-linear and intricate relationships among variables, which is essential for precise predictions in survival analysis. Techniques such as random survival forests are particularly effective in managing right-censored data often found in survival analysis (Suresh et al., 2022). Given these insights, we aimed to explore predictions from a non-linear perspective. We next evaluated the Naive Bayesian model, which appeared promising due to its appropriateness for our scale of analysis. However, the extensive size of our dataset exceeded our computational capabilities, precluding a comprehensive application of this model. Consequently, we turned to the Random Forest model. The Random Forest algorithm was selected to mitigate the challenges of data overlap and extensive computation times encountered with the SVM and Naive Bayesian models. Analyzing the decision trees generated by this model yielded substantial insights into the factors influencing business survival rates. For instance, it was found that businesses with fewer than 0.5 nearby parking spaces situated in areas where population density does not exceed 5.5 thousand per square mile generally fall into a 'medium' survival class. Additionally, a competitor count of fewer than 47.5 was associated with businesses predominantly in the 'medium' survival category. These findings underscore the benefits of minimal competition and adequate parking on business survival, corroborating earlier observations regarding the significance of population density, parking availability, and the competitive landscape.

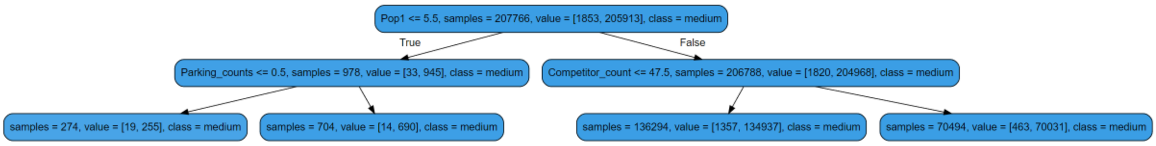
Pop1 <= 5.5, samples = 207766, value = [1853, 205913], class = medium
True
False
Parking_counts <= 0.5, samples = 978, value = [33, 945], class = medium
Competitor_count <= 47.5, samples = 206788, value = [1820, 204968], class = medium
samples = 274, value = [19, 255], class = medium
samples = 704, value = [14, 690], class = medium
samples = 136294, value = [1357, 134937], class = medium
samples = 70494, value = [463, 70031], class = medium

Figure 3. Random Forest Visualization

Discussion.

Our model was developed to empower stakeholders and startup owners with actionable insights for establishing businesses in New York City. We identified key factors influencing business survival by evaluating the demographic characteristics across different geographical locations. Utilizing the Random Forest algorithm, our findings highlighted that population density significantly affects business success. Moreover, the presence of competing businesses in close proximity also emerged as a critical factor. Interestingly, among accessibility metrics, the availability of nearby parking facilities was the sole factor with a significant impact on business survival rates. These insights are invaluable for prospective business owners, underscoring the critical importance of location demographics and accessibility within the urban landscape of NYC.

Additionally, the integration of machine learning models such as Random Forest, SVM, and Naive Bayes in our analysis highlights the potential of these technologies to provide nuanced insights into urban business dynamics. The Random Forest model, in particular, offered valuable predictions by effectively handling non-linear relationships and large datasets, thereby identifying critical factors influencing business survival. For instance, our findings emphasize the significance of parking availability and competitive landscape, which are often overlooked in traditional analyses. These insights not only underscore the multifaceted nature of urban economic environments but also provide a robust framework for urban planners and policymakers to devise targeted strategies that enhance business viability. By leveraging machine learning techniques, future research can further refine these predictive models to accommodate additional variables and improve the accuracy of business survival predictions. This approach represents a significant advancement in urban economic studies, paving the way for more data-driven and effective policy interventions that can adapt to the rapidly changing dynamics of urban centers.

Conclusions.

The comprehensive analysis of this study confirms the critical impact of economic factors, population dynamics, and accessibility on business viability and success in urban environments, particularly in New York City. The findings suggest that median household income, population density, and accessibility are key determinants of commercial viability. Higher median household incomes are associated with increased consumer purchasing power, promoting local business activity, especially in higher-income neighborhoods that support diverse business operations and services. Therefore, increasing overall income levels in the region may be an effective strategy for promoting business prosperity. Areas with higher population densities provide a larger customer base, critical to business prosperity, but also lead to more intense competition and increased operating costs, requiring businesses to balance these challenges and opportunities by offering differentiated goods and services to attract a diverse customer base. Accessibility is crucial to business success in urban environments; proximity to transportation hubs like subway stations and bus stops improves accessibility for both customers and employees, enhancing commercial performance. In less accessible areas, providing adequate parking facilities can significantly influence customer choice and commercial viability. The policy environment and local government economic support measures play a decisive role in business success. Policymakers should implement incentives for entrepreneurship, such as tax incentives, loan support, and business training, to increase the survival and success of small businesses. The study emphasizes the importance of supportive economic policies and infrastructure development that enhance accessibility and demographic advantages within urban areas, which urban planners and policymakers should consider when designing strategies to promote business growth and sustainability. Future research should explore avenues to further understand the dynamics of business survival, building on the findings of this project.

References

Bureau, U. C. (2023). American Community Survey Data. Retrieved from https://www.census.gov/programs-surveys/acs/data.html

Department of Consumer and Worker Protection (DCWP). (2024). License applications: NYC open data.

Furman Center for Real Estate and Urban Policy, New York University. (n.d.). Interactive data visualization tool for median household income adjusted by Census Tract in New York City, 2014-2018.

NYC Planimetric Database: Parking lot: NYC open data. (n.d.). Retrieved from https://data.cityofnewyork.us/City-Government/NYC-Planimetric-Database-Parking-Lot/h7zy-iq3d

Pathak, S., Quraishi, S. J., Singh, A., Singh, M., Arora, K., & Ather, D. (2023). A comparative analysis of machine learning models: SVM, Naïve Bayes, Random Forest, and LSTM in predictive analytics. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS). https://doi.org/10.1109/ictacs59847.2023.10390255

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and Research Directions. SN Computer Science, 2(3), Article 00592. https://doi.org/10.1007/s42979-021-00592-x

United States Census Bureau. (2020). 2020 Census Data: Basic demographic and housing characteristics for New York City’s boroughs, community districts, city council districts, neighborhood tabulation areas, and census tracts.

Suresh, K., Severn, C. & Ghosh, D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol 22, 207 (2022). https://doi.org/10.1186/s12874-022-01679-6

Appendix

Contribution:

Qingyuan Feng(30%) : Data cleaning data visualization logistic regression Naïve Bayes analysis report writing, poster writing.

Mingchuan Ma(30%) : Poster Design, Random Forest, Data Cleaning, Report writing and format.

Xianze Gao(20%) : Data collection, Report writing, literature review ,poster writing.

Jiuling Zhong(20%) : Data collection, Report writing, literature review, poster writing.