Business Survival Rate in NYC: Focusing on Median household Income, Population and Accessibility

Mingchuan Ma, Qingyuan Feng, Xianze Gao, Jiuling Zhong

NYU MLC Spring 2024

Introduction

In the dynamic field of economics, the robustness of businesses within a city is indicative of its overall economic health and long-term sustainability. Our project leverages multiple data sources to construct a model that predicts the operational status of businesses in New York City. This initiative serves not only to evaluate the current business environment but also to inform future urban planning and policy-making. Utilizing an interdisciplinary approach, we analyze data from the 2020 Census, business license applications, established business records, median household income figures, and transportation infrastructure. This comprehensive analysis aims to identify the key factors influencing business operations across different areas.

Research Question:

To what extent does Median Household Income, Population, and Accessibility affect the survival rate of business in New York City

Methods

The first phase of the study processed the data by (1) compiling data on New York City businesses, license applications, subway stations, bus stops, and parking lots. The data was obtained from NYC Open Data, NYC Furman Center. (2) Population data and median income data were broken down by census tract and a geographic map of median population input was created. (3) Integrate the data to determine the survival of businesses and the impact of different transportation facilities on business activity in each location.

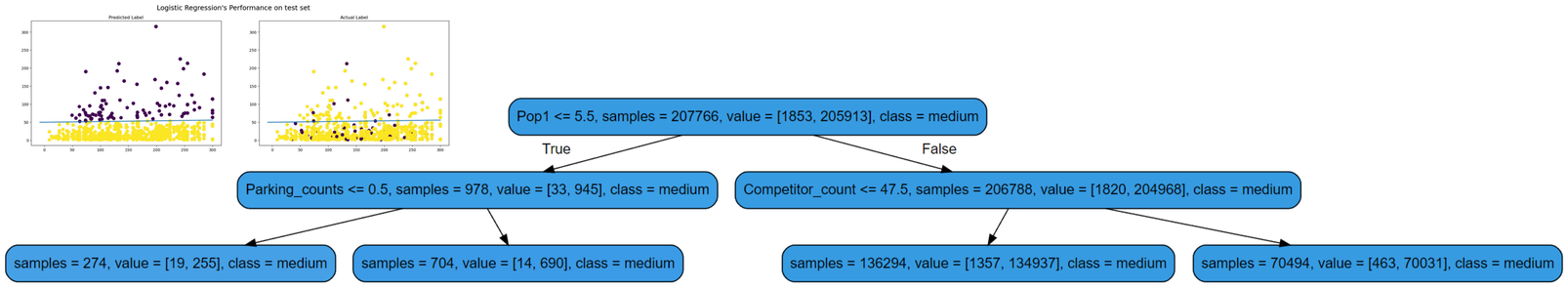
In the second phase, we have prepared two visualizations.The first image visualizes the relationship between population distribution as well as industry distribution in New York City. The second station map visualizes the relationship between median income and industry distribution of the population across New York City.

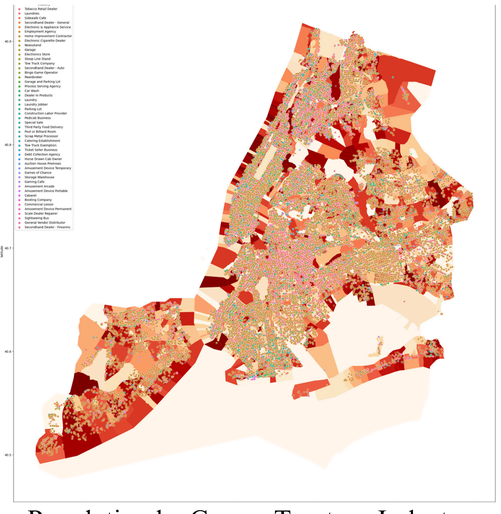
In the third phase, we trained Random Forest, Support Vector Machine (SVM), and Naïve Bayes to unravel the connections that shape urban commercial environments. The datasets cover a range of factors from information to accessibility measures like proximity, to public transportation hubs and parking availability.

Results

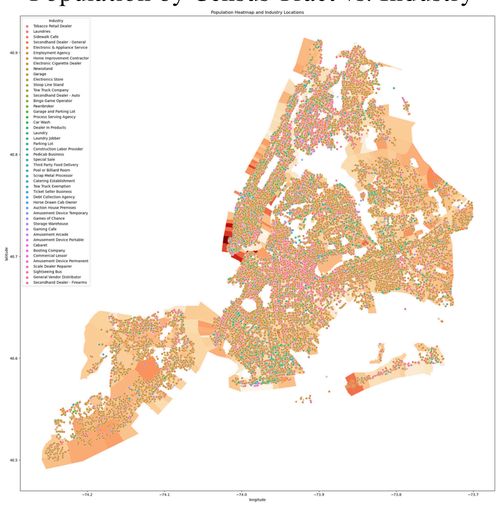
Population data and types of businesses were incorporated as input variables into the SVM (Support Vector Machine) model. However, a significant discrepancy was observed between the predicted outcomes and the actual business statuses, which were labeled as active or inactive. The extensive overlap between these labels rendered the SVM model ineffective for predictive purposes. Similarly, the Naive Bayes model proved impractical for the analysis; it required over two hours to process and failed to produce any conclusive results.

To circumvent the difficulties encountered with the SVM and Naive Bayes models, the Random Forest model was utilized. This approach effectively addressed the challenges of overlapping labels and long processing times, identifying population density, the number of parking spaces, and the presence of competitors as significant factors influencing business survival.





Population by Census Tract vs. Industry



Median Income by Census Tract vs. Industry

Discussion

The model we developed aims to assist stakeholders and startup owners make informed decisions when establishing businesses in New York City. We can identify critical factors that influence business survival by assessing the demographic characteristics of various geographic locations. Our analysis, conducted using the Random Forest algorithm, indicates that the population density of an area is a significant determinant of business success. Additionally, the number of competing businesses in the vicinity also plays a crucial role. However, concerning accessibility, the availability of nearby parking facilities is the only factor that significantly impacts the survival rate of businesses. This model provides valuable insights for prospective business owners on the importance of location demographics and accessibility in the urban landscape of NYC.

References Data sources

- Bureau, U. C. (2023). American Community Survey Data. Retrieved from https://www.census.gov/programs-surveys/acs/data.html
- Bus stop shelters. (n.d.-a). Retrieved from https://data.cityofnewyork.us/Transportation/Bus-Stop-Shelters/qafz-7myz
- Department of Consumer and Worker Protection (DCWP). (2024). License applications: NYC open data.
- Department of Consumer and Worker Protection (DCWP). (2024). Legally operating businesses: NYC Open Data. Retrieved from https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh/about_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.
- Department of Health and Mental Hygiene (DOHMH). (2020). Modified ZIP code tabulation areas (MODZCTA): NYC open data.
- 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). doi:10.1109/ictacs59847.2023.10390255
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and Research Directions. SN Computer Science, 2(3). doi:10.1007/s42979-021-00592-x
- Subway lines. (n.d.-b). Retrieved from https://data.cityofnewyork.us/Transportation/Subway-Lines/3qz8-muuu
- 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.