Team 105 Final Report

Idris Kuti, Gautam Matta, Akash Nikam, Junfei Xia, Chris Walker

November 2022

1 Introduction and Motivations

Data analytics and visualization is a task carried out by large firms with significant resources. This is due in part to the monetary and time investments required for a firm to successfully leverage analytics in their decision-making. Small business owners often lack the resources to use data in a similar fashion to large firms. Our project aims at reducing the analytics gap between large and small firms.

One of the most important decisions a small business needs to make is the location of their storefront or office. We developed an interactive GUI application called StratusHere which helps small business owners determine the viability of a given location for their business. To build this application we collected data, created models, and developed data visualizations. Our primary objectives were to:

a. build a classification model to predict the business category of commercial properties,
b. incorporate this model into a map GUI on a map similar to Zillow or Redfin,
c. and show a heat map of suitable regions for conducting business in Atlanta.

2 Problem Definition

Small businesses do not have the resources to perform complex analytics to help them decide where to open or expand their business. While business owners may work with a commercial property agent, they often lack the information necessary to make a data driven decision. Therefore, small business owners often compete with large firms for the best locations. We built an application to address this problem.

3 Literature Survey

Large firms and academic experts use complex modeling techniques to estimate an empirical event (number of customers per month, for example) using information based on location. The upfront cost of data collection and the associated analytics is high for even very profitable and successful firms. For example, Karande & Lombard (2005)[13] made an empirical investigation for location strategies of broad-line retailers. According to research by Shuihua et al. using a machine learning-based model for selection of a business location using factors like sales potential based on neighboring businesses we can eliminate the current subjective approach to selecting a location for a business[1]. Kuo et al. proposed a four components decision-making system for convenience store location selection[12]. Rosa et al. (2022) use network theory to analyze the commercial

spatial interactions[14]. One of the methods discussed by Jensen et al. (2014) uses temporal features by making probabilistic assumptions in the data to combine it with GIS data and create visualizations[7]. Alan Murray discusses the replicability challenges in location analysis, especially the concerns in spacial optimization and unreliability of the heuristic methods[8].

In addition, research conducted by JP Morgan Chase suggests there is a correlation between low-profit margin small businesses and the lack of high-tech or other professional services firms in certain locations[3]. This finding was useful for our project because it gives us a reference point for selecting possible locations for small businesses in relation to professional service firms. This suggests that without these types of services readily available to small businesses, these businesses are struggling to optimize their locations. Our product will help bridge the gap between high-tech firms, academia, and small businesses, leveling the playing field without the investment in the same resources. We will also eliminate some of the risks of picking a bad location to start a business as well as reduce the lead time required to go to market therefore increasing the vibrancy of the community and enriching the economy.

Our tool runs the risk of misleading a user by guiding them to a poor decision which has a monetary cost. Our data also has implicit biases present from the real world business environment. The paper on local context and neighborhood conditions explains understanding the uniqueness of certain regions, their spacial bounded characteristics, and how they affect the location making decisions. This research helps in contextualizing the importance of economic, demographic and geographical conditions at a neighborhood level helping us reduce bias[2]. According to the empirical investigation by Kritzinger et al, location had the weakest positive relationship with business performance which is contradictory to other literature findings[11]. Instead, it was discovered that macro-environmental factors such as employment, green buildings, rental rates, inflation had positive relationship with business performance.

Considering all of these factors, our tool is aimed at providing a relative rank ordering of business locations for small business owners. Providing a rank ordering, rather than a mutually-exclusive classification, allows the business owner to merge their intuition with model results. We also will not arbitrarily remove recommended properties from a user's view and will simply display an Okay rating for the worst properties. Conversely, other properties will earn an Good or Great ranking.

4 Proposed Methods

Data Collection & Data Cleaning

We designed a model structure that was conducive to available data. Ideally, we want to predict whether a given property was expected to be successful for a given business category. However, this is difficult to quantify. Instead, we decided that if we could predict the business category (i.e., grocery store, restaurant) of an existing business in the Atlanta metro area, then we could apply new properties to the model. We can use these predictions to indicate whether a property is Okay, Good, or Great for a given business type.

With these constraints in mind, we utilized a business database provided by Georgia Tech called Mergent Intellect. This database includes fields like latitude/longitude, zip code, business sales, business category, and more. We collected data on 150k retail businesses in the Atlanta metro.

These businesses must have less than 100 employees at a given location as our project is oriented towards small and medium sized firms. Data could only be downloaded in 20k record chunks. We downloaded 6 individual CSV files and combined them. We have also downloaded traffic data from Georgia Traffic Monitoring Program and interpolated the data with Gaussian Progress Regression and KNN for each business location.

Alongside business-level attributes, we also collected census-level data including population, income, etc at the zip code level using Georgia Tech provided business database called Simply Analytics. This data was left joined to our business-level data to create a final data set with 150k records and approximately 20 attributes.

An important step before any modeling can be done is to explore the data and make sure we are getting the most useful information and no information is missing. When we pulled the data there were a lot of columns which contained information that could have been useful but after some analysis and some basic statistics like a correlation matrix, we were able to select only the features that help predict our response variable. Other fields like distribution of categories, square footage against zip codes, distribution of sales numbers was also conducted but due to time constraints not all of these features made it into our MVP. While smaller than some other data sets, this data set was large enough for our use case. In addition, we will discuss some big-data innovations needed for our web application in later in this document.

Model Design & Web Application

Based on the 150k records of small business information, we trained a classification model for different categories. RandomForestClassifier is selected as the algorithm because of the accuracy of prediction, size, and speed of the model when hosted on the application platform.

Our web application is an R Shiny application hosted on shinyapps.io. Shiny is a powerful web framework that allows for rapid prototyping of robust data-driven applications. While there are alternatives for Python (streamlit apps, and an upcoming version of Shiny for Python) we felt this was the best framework for our use case despite model estimation occurring in Python.

The reticulate package allows Python virtual environments to be run within R. At first, we planned on using this package to load our random forest as a pickle file and serve it as part of our application. However, reticulate is slow to launch a VM and created a bad experience for users. Instead, we exported our random forest in a text format similar to JSON. This text was parsed into nested-lists in R. We developed a C++ module which runs underneath the R Shiny app using Rcpp (a C++ API for R). Our random forest is predicted recursively in C++. Because C++ is a compiled language (unlike R and Python which are interpreted), we can generate hundreds of forecasts reliably and at great speed. Given the complexity of our random forest, we are computing over 10 million recursive computations each time a user makes a change in under 0.2 seconds. A similar operation in R or Python would take 1-2 minutes creating an unacceptable experience for our users. Our web app is served with application code, the C++ module, and a serialized random forest.

Our web app is called StratusHere. The web application is loaded and users are presented with a few options on the left-hand side. This allows users to select their business category, a zip code of interest, and input an estimate of their annual sales. Our model then applies these attributes to property data which already contains zip-level attributes.

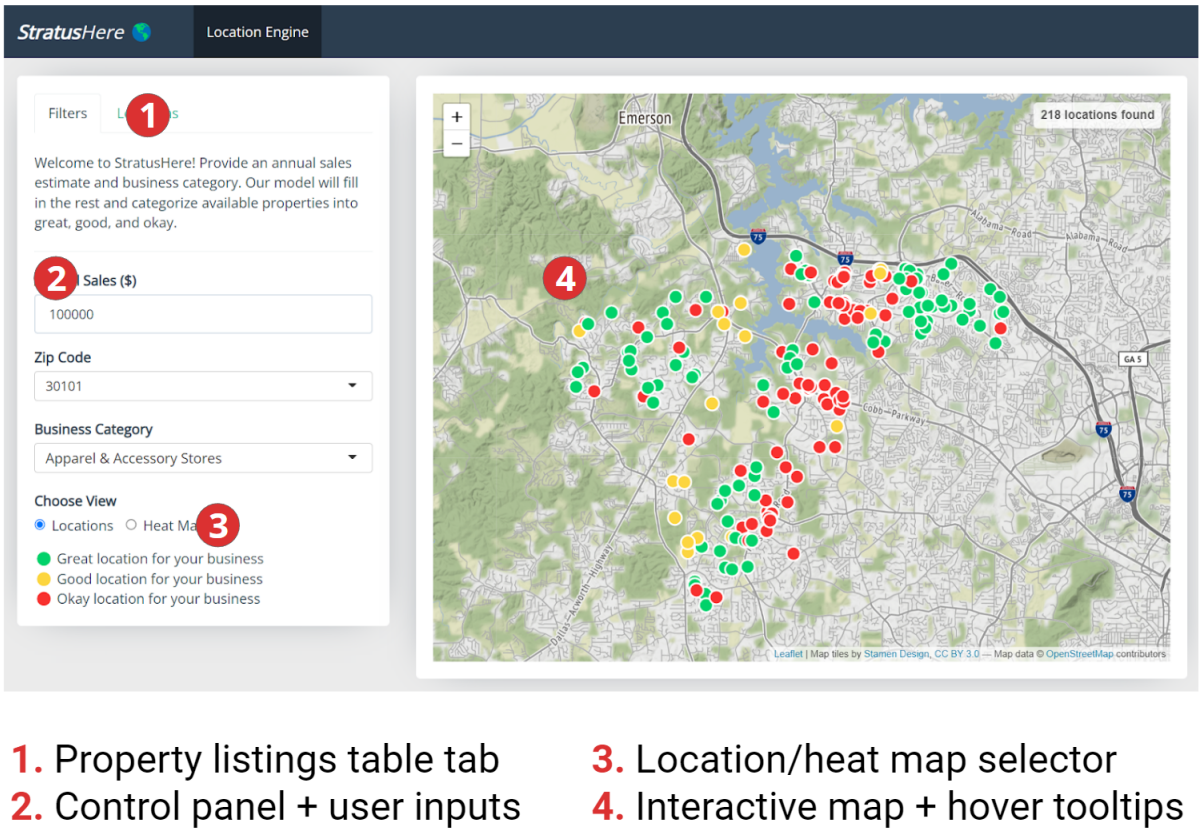

Figure 1: This is how the webpage looks like.

Each property is passed through our random forest which contains 100 trees. Each tree within the forest produces a business category classification. For a given property, we tabulate the number of trees which predict a given category. Suppose we have the following tree tabulations for three properties and three business categories:

Property	A	B	C
1	30	60	10
2	20	70	10
3	10	80	10

If a user is looking for recommendations in business category A, we look at the number of trees predicting business category A. From there we find the median number of trees. In this case, 20 is the median. Therefore, property 1 is green because it is above median whereas property 2 is at the median and is colored yellow. Finally, property 3 is below median so it is red. If the user switches to a different category, then tabulations would be used for a different tree outcome.

This red, yellow, green framework protects StratusHere from arbitrarily removing properties. In addition, we are better shielded from making a poor recommendation because no properties are explicitly filtered out, we simply suggest which properties may be a better fit. We also give users the ability to view a heat map. The heat map generates blobs from properties which are colored yellow or green.

5 Experiments & Evaluation

The first step was to gather data used to train our models. To simplify the initial process so we can get a minimum viable product available in such a short time frame, we limited the search space to the Atlanta Metro. With that, we were able to gather a list of 150k businesses in the Atlanta region with associated attributes. With this data, some data cleaning was performed so we could use it to build a classification model. This data was merged with demographic data to create our final modeling data set.

Machine Learning Methods

This is a classification problem with the type of business category as target variable and location and sales based inputs as features. The sampling of data was done such that all business categories are represented in the data to avoid any skewed bias. scikit learn library is used for all the machine learning operations. One hot encoding is used to convert the categorical variables like zip code, minority owned, etc into numerical. Data is split into training / testing in a 80-20 split. GridsearchCV is used to tune the hyper parameters in all of the models mentioned below before presenting the best results.
After experimenting with various classification algorithms like XGboost classifier, K Nearest Neighbors, Decision Tree classifier, etc. Random Forest Classifier was finalized to be used in the application. Here are the results of experimentation with some of the machine learning models.

Algorithm	**Accuracy for top result**	**Accuracy for prediction to be in top 3**
RandomForest	0.328	0.659
XGBoost	0.27	0.634
Decision Tree	0.184	0.291
KNN	0.274	0.60

6 Conclusions & Discussion

Choosing a location for an office or storefront is one of the most crucial choices a small business must make and they do not have resources to perform the analytics when making such decisions. As a part of our project initiative, we have built an interactive GUI application which aids small business owners in assessing the suitability of a potential location for their operations.

The major objectives were to gather and cleanse the data, experiment and build a classification model to forecast business categories, incorporate this into GUI and show a heat map of potential business areas in Atlanta metro region. We finalized the Random Forest Classifier after experimenting with various other classification models mentioned earlier and are satisfied with the range of results it has provided.

Currently, we have built the application keeping Atlanta in mind. We can expand this to other metro regions and scrape data from other metros. We can also scrape more features to improve our model's accuracy. Adding some more filters such as square footage, hours of operations, which we couldn't do earlier because of time constraints, would be a great idea for the future. The application is currently hosted on shinyapps.io with about 100 active hours and we can obtain more credits if we need to scale it up for a wider audience.

Our group was able to overcome challenges related to programming languages (C++ over Python for faster results), web hosting (switched to Starter from Free tier of shinyapps.io), data collection (scraping, categorizing various sub-categories, limited data etc) and more.

Overall, we are proud of the data, model, and application we were able to construct within one semester. All group members made meaningful contributions to our project. Here is the distribution for team members' efforts:

Activity	**Owner**
Data Gathering and EDA	Idris, Chris, Akash, Gautam, Junfei
Data Cleaning	Idris, Chris, Akash, Junfei
Modeling	Akash, Gautam, Junfei
GUI Building	Chris, Akash, Gautam
Documentation	Junfei, Akash, Idris, Chris, Gautam

7 References

[1]Han, S., Jia, X., Chen, X., Gupta, S., Kumar, A., & Lin, Z. (2022). Search well and be wise: A machine learning approach to search for a profitable location. Journal of Business Research, 144, 416-427.
[2]Sefiani, Y., Davies, B., & Bown, R. (2016). The perceptual effects of location on the performance of small businesses.
[3]Farrell, D., Wheat, C., & Grandet, C. (2019). Place matters: Small business financial health in urban communities. Available at SSRN 3462771.
[4]Kuo, R. J., Chi, S. C., & Kao, S. S. (2002). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. Computers in industry, 47(2), 199-214.
[5]Marinković, S., Nikolić, I., & Rakićević, J. (2018). Selecting location for a new business unit in ICT industry. Zbornik Radova Ekonomski Fakultet u Rijeka, 36(2), 801-825.
[6]Goodchild, M. F. (2005). GIS and modeling overview. GIS, spatial analysis, and modeling. ESRI Press, Redlands, 1-18.
[7]Jensen, M., Gutierrez, J., & Pedersen, J. (2014). Location intelligence application in digital data activity dimensioning in smart cities. Procedia Computer Science, 36, 418-424.
[8]Murray, A. T. (n.d.). Replicability Challenges in Location Analytics. Home. Retrieved October 12, 2022, from https://scholarspace.manoa.hawaii.edu/
[9]Al Sonosy, O., Rady, S., Badr, N. L., & Hashem, M. (2016, December). A study of spatial machine learning for business behavior prediction in location based social networks. In 2016 11th International Conference on Computer Engineering & Systems (ICCES) (pp. 266-272). IEEE.
[10]Schmidt , P. J., Feng, J., & Freeze, R. (2021, January 5). Geographic Data Informs Funding and Management of Metro Bike Share System. ScholarSpace. Retrieved October 12, 2022, from https://scholarspace.manoa.hawaii.edu/
[11]Barnard, S., Kritzinger, B., & Kruger, J. (2011). Location decision strategies for improving SMME business performance. Acta Commercii, 11(1), 111-128.
[12]Kuo, R. J., Chi, S. C., & Kao, S. S. (2002). A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. Computers in industry, 47(2), 199-214.
[13]Karande, K., & Lombard, J. R. (2005). Location strategies of broad-line retailers: an empirical investigation. Journal of Business Research, 58(5), 687-695.
[14]Sanchez-Saiz, R. M., Ahedo, V., Santos, J. I., Gómez, S., & Galán, J. M. (2022). Identification of robust retailing location patterns with complex network approaches. Complex & Intelligent Systems, 8(1), 83-106.
[15]AlSonosy, O., Rady, S., Badr, N., Hashem, M. (2017). Business Behavior Predictions Using Location Based Social Networks in Smart Cities. In Information Innovation Technology in Smart Cities (pp. 105–122). Springer Singapore. https://doi.org/10.1007/978-981-10-1741-4_8