RISK ANALYSIS OF SETTING UP A RESTAURANT AT NYC

Santoshi Laxmi Reddy Ellanki[1] and John Jenq[2]

[1]University of Texas Medical Branch, Houston, Texas, USA
[2]Department of Computer Science, Montclair State University, NJ, USA

ABSTRACT

In this report, a system was developed that can predict the outcome of opening a restaurant in NYC based on various NYC open data sets, such as 311 calls, New York Police crime records and restaurant rating data. The data sets were preprocessed and cleaned before analysis to improve the quality of our results.

KEYWORDS

Big data, Risk analysis, PySpark, Decision tree.

1. INTRODUCTION

In today's highly competitive world, every business has a motive to be profitable. Among business sectors, restaurant's business is more connected to locality they are set up in. Apart from the quality of the food provided in the restaurant, there are other factors that need to be considered in order to make a restaurant business successful. People often choose restaurants that are in a safe and secure locations so they can relax and enjoy the meal. Some of the most important factors to consider regarding safety and accessibility are crime rate, entertainment, ease of commute and infrastructure. For example, an area with higher average income, and lower complaints of rodents, potholes, etc. can be considered as more safe and secure than other areas.

At an estimated population of 8.4 million, New York City is the most populated city in the United States. There is a constant fluctuation in the number of people moving to and from NYC every year, which has led to it having one of the most dynamic real estate markets worldwide. With such dynamic nature which will affect the locality rating, there is a need to constantly analyse the different localities of NYC. Someone moving into the city for setting up a business (e.g., restaurant) might be very interested in knowing which areas are more likely to bring in larger profit, which areas have good amenities, and less crime complaints. In this regard, metrics that show how safe a locality is, or which localities are prone to higher crime rate or higher service requests, are important factors to consider before setting up a restaurant. Financial institutes need to analyse risk when a business owner want to get finance from them. A tool to analyse risk would be beneficial.

Customer ratings for a restaurant are one of the important factors which contribute to the restaurant's success. If we take into consideration crime reports, 311 requests and the ratings of other restaurants in the same locality, the success of opening a new restaurant in that area can be predicted. These of course depend on the quality of the restaurant itself and how its facilities are. But we believe that the surroundings do affect the success of a restaurant. This is where the analytic comes in.

David C. Wyld et al. (Eds): CCSIT, SIPP, PDCTA, AISC, NLPCL, BIGML, NCWMC - 2021
pp. 21-27, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110703

In this report, our goal is to come up with a dynamic model which can predict the outcome of opening a restaurant in NYC. The same can be applied to other cities in the US. To achieve this, we need to find relevant data and perform analysis on that data to define a dynamic and robust model. The advent of Big Data and advancements in computational techniques have enabled us to optimally find metrics that can help us to choose big data. We want the data set to be as substantial and complete as possible so that it covers a long history of transactions and information. The more comprehensive the data, the bigger the possibility of building an accurate model and producing meaningful results. The data sets we are using are from NYC open data website. This website will be used to import data sets for 311, NYPD and restaurant data. The data sets were comprehensive enough to suit our needs. [6, 7, 8].

Since there are only few technologies that can handle huge data, it is important that we choose a technology which is freeware or processed with a minimal cost. When processing large data, one machine with more memory and disk space is more expensive and less efficient than a group of machines with cheap hardware and configuration. As a result, we turned to big data technologies to make use of cluster machines and lightning-fast processing of data. After deciding to use big data technology to process the data, it is also important to choose the correct big data technology. We will choose a popular technology, Spark [5], which is built on Hadoop [4].

As Hadoop is the heart and soul of big data technologies to create and maintain a cluster, it is important to take advantage of Hadoop for cluster management and HDFS. We also need a technology to perform data operations, so we need to choose a platform that can handle huge data and perform operations. As Hive [11] is a popular data warehouse technology, we can use Hive for any data operations. These big data technologies are able to complete our data processing operations an efficient manner, so we decided to use them.

During the planning stages, in order for restaurant owners to accurately predict the success of opening a new restaurant, a few important factors need to be considered. These predictions would help the restaurant owner decide where to set up a restaurant based on location, crime and other factors to reduce the business risk. To make this prediction, we first need to identify patterns in the data so that the model can learn from these patterns and apply them to predict future behaviours and patterns. There are many algorithms that we can use to do this, and in this report, we used the simple decision tree algorithm [9, 12] because its similarity to human thinking and ease of use both result in good interpretations of the data. Some researchers even use this approach to identify risk factors for relapse to Smoking [2]. If we are clear on which tree nodes and attributes to choose, we can definitely produce a dynamic and robust model. Decision tree is a classification technique and the decision tree algorithm tries to solve the problem by using a tree representation. Each internal node of the tree corresponds to an attribute and each leaf node corresponds to a class label.

2. SYSTEM IMPLEMENTATION

Ubuntu [10] was chosen as the platform and Spark was installed on top of it. The data source includes 10 gigabytes of 311 service requests, where each record includes the type of call, latitude and longitude of the incident, zip code of the incident and the date of the complaint. The second data set is 1.4 gigabytes and includes New York City Police Department records reported crime and offense data based upon New York State Penal Law and other New York State Laws. Records specify type of crime, latitude and longitude of the incident, zip code of the crime scene and the date of the crime complaint. The third data set is 300 MB of restaurant rating records which includes restaurant zip code, building, street, and restaurant rating. Figure 1 shows the system components and their connections.

For these three major datasets, each contained millions of records. A substantial amount of time was spent on data processing and preparing the data sets for analysis. The data sets were cleaned using Hive SQL, where we handled missing values, duplicate values, records with invalid zip codes, records with invalid latitude and longitude information and determined the important fields to retain.

Some of the records were recorded at a latitude and longitude level instead of at a zip code level. Thus, we used the Geocode API and developed a python script that accepts latitude and longitude information as input parameters and outputs the corresponding zip code. Using Hive SQL, we joined the zip code of respective latitude and longitude location with the remaining attributes in the dataset. Finally, we used SQL and Hive to further process the fields which had a multitude of information contained in lists. These steps were taken in order to prepare individual datasets for the desired analytics. These refined individual datasets were grouped together using Hive SQL and SQL.

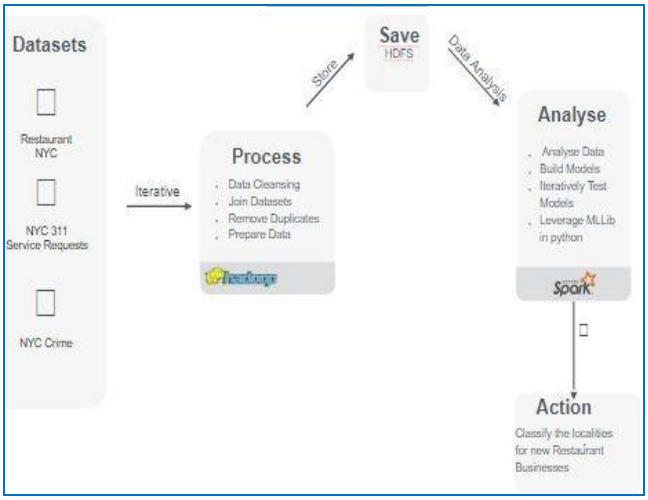

Figure 1. System components

2.1. Severity Classification Based on 311 Requests

We used 311 requests from past 5 years. Features that we considered after cleaning the data sets are 311_Incident Zip, which represent the 228 zip codes of NYC, and 311_Severity_1, 311_Severity_2 and 311_severity_3, which represent the count of 311 complaints based on severity. Complaints such as food poisoning and drug activity … etc., come under 311_Severity_1. Sidewalk condition, curb condition, mosquitoes …etc., come under 311_Severity_2 and complaints such as no permit parking, or lack of public bathroom come under 311_Severity_3. For each zip code, we calculated the count of all the severities and pivoted the severity counts as 311_Severity_1, 311_Severity_2, 311_Severity_3 for each zip code using SQL. See figure 2 for the classification.

311 Complaint severity	311 Complaint types
311_Severity_1	Food poisoning, Drug activity etc.
311_Severity_2	Sidewalk condition, curb condition, mosquitoes
311_Severity_3	No permit, public toilet

Figure 2. Classification of 311 Severity Based on 311 Complaint Types.

2.2. Severity Classification Based on Crime Complaints

We used crime complaints from past 5 years. Features that are considered after cleaning the data sets are Crime_Incident_Zip, which represent the 228 zip codes of NYC, and Crime_Severity_1, Crime_Severity_2 and Crime_severity_3, which represent the count of crime complaints based on severity. Crimes such as murder, felony … etc., come under Crime_severity_1. Forgery, robbery, assault … etc., come under Crime_severity_2 and the miscellaneous complaints come under Crime_severity_3. Similar to the 311 severity data sets, we calculated the count for each zip code with all the severities and pivoted the severity counts as Crime_Severity_1, Crime_Severity_2, and Crime_Severity_3 for each zip code using SQL. See Figure 3.

Crime Complaint severity	Crime Complaint types
Crime_Severity_1	Felony, murder, etc.
Crime_Severity_2	Assault, robbery, forgery
Crime_Severity_3	Miscellaneous

Figure 3. Classification of crime severity based on crime complaint types

2.3. Determine Restaurant Label Based on the Restaurant Rating

The features considered in the restaurant dataset are the Restaurant zip of the NYC locality, restaurant address, type of cuisine and the rating of the restaurant. The average of the ratings given to all the restaurants are considered. Restaurant rating which 2.5 or below is labelled as “low”, which is low recommended restaurant, Restaurant rating greater than 2.5 and less than 3.3 is labelled as “medium”, which is a medium recommended restaurant. Restaurant rating greater than 3.3 is labelled as “high”, which is highly recommended restaurant. See Figure 4 for restaurant labelling

Restaurant Label	Restaurant Rating
Low	2.5 or less
Medium	>2.5 and <3.3
High	3.3 and above

Figure 4. Restaurant Labelling Based on Rating

2.4. Determine the 311 Status, Crime Status Based on Crime Complaint Severity and 311 Complaint Severity

If the 311_Severity_1 is greater than 311_Severity_2 and 311_Severity_3, then 311_status is labelled as “low” and therefore is low recommended location. If the 311_Severity_2 is greater than 311_Severity_1 and 311_Severity_3 then the 311_status is labelled as “medium”, which is medium recommended location. Similarly, if the 311_Severity_3 is greater than 311_Severity_1 and 311_Severity_2 then the 311_status is labelled as “high” and this means the area is highly recommended. See Figure 5 for 311_status classification.

We classified crime severity in a similar way. If the Crime_Severity_1 is greater than both Crime_Severiy_2 and Crime_Severity_3, then Crime status is marked as “low” which means it is not a recommended location. If the Crime_Severity_2 is greater than both Crime_Severity_1 and Crime_Severity_3, then Crime status is labelled as “medium” and it is medium recommended location. If 311_Severity_3 is greater than both 311_Severity_2 and 311_Severity_1, then the

Crime status is labelled as “high” and it is a highly recommended location. Figure 6 shows the classification of crime status based on the crime complaint severity.

311 Complaint severity	311 status
311_Severity_1>311_Severity_2 & 311_Severity_1>311_Severity_3	Low
311_Severity_2>311_Severity_1 & 311_Severity_2>311_Severity_3	Medium
311_Severity_3 > 311_Severity_1 & 311_Severity_3 > 311_Severity_2	High

Figure 5. The 311 Status classification

Crime Complaint severity	Crime status
Crime_Severity_1 > Crime_Severity_2 & Crime_Severity_1 > Crime_Severity_3	Low
Crime_Severity_2 > Crime_Severity_1 & Crime_Severity_2 > Crime_Severity_3	Medium
Crime_Severity_3 > Crime_Severity_1 & Crime_Severity_3 > Crime_Severity_2	High

Figure 6. Crime Status classification

2.5. Determine the Location Label Based on Crime Status, 311 Status and Restaurant rating

Looking at all the possible combinations of crime status, 311 status and restaurant rating, we labelled each area as best, moderate or worst. For instance, if the 311 status is low, crime status has a low classification and restaurant rating is less than 2.5, then we label this location as worst, which means the location is not recommended for setting up a restaurant. See Figure 7 for classifications.

311 status	Crime status	Restaurant Rating	Location label
Low	Low	<2.5	Worst
Low	Medium	<2.5	Worst
Medium	Low	<2.5	Worst
Medium	Medium	<2.5	Worst
High	Low	<2.5	Worst
High	Medium	<2.5	Worst
High	Low	2.5-3.3	Worst
Medium	Low	2.5-3.3	Worst
Low	Low	2.5-3.3	Worst
Medium	Medium	2.5-3.3	Moderate
Low	Low	>=3.3	Moderate
Medium	Low	>=3.3	Moderate
Low	Medium	>=3.3	Moderate
High	Medium	2.5-3.3	Moderate
High	Medium	>=3.3	Moderate
High	High	>=3.3	Best
High	Low	>=3.3	Best
High	Medium	>=3.3	Best
High	Medium	>=3.3	Best

Figure 7. Location Classification

2.6. Prediction Model Implementation

The Apache Spark MlLib was used to develop the prediction models. The feature set was as described above. We experimented with two different Machine Learning algorithms: Logistic Regression and Decision Trees. Logistic Regression is best suited for models with output ranging from 0 to 1. In this model, output was categorized as low, medium, or high ranging from 0 to 2. The dataset obtained after cleaning is split randomly by using the random function into both training and the testing datasets, known as K-fold cross validation approach. The model after training the dataset is applied to predict the results for the testing and the results are evaluated. We found that compared to decision trees, Logistic regression was a less accurate algorithm.

3. CONCLUSIONS AND REMARKS

The total count for prediction with low recommended is 4149036. For medium recommended, total count is 3984406 and for high recommended, 3273156. See Figure 8 for a pie chart illustrate the percentage of the recommendation. Due to the communication delays among cluster PCs the speed up is not significant.

Decision Tree Algorithm is used as a classification technique which is used on the existing dataset and the resulting decision tree is used to create a strategy for finding the status of the zip code based on the security and severity. The decision tree considers the factors that are considered relevant for the decision.

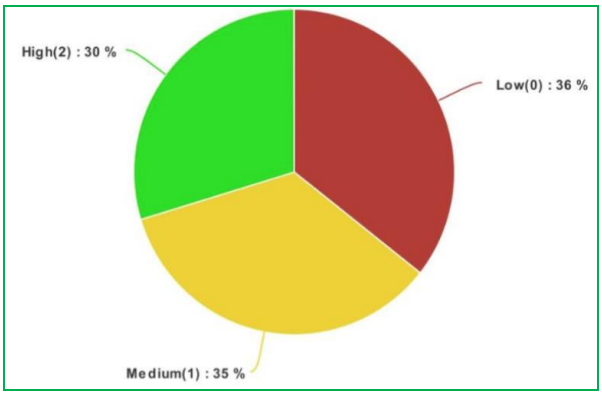

Figure 8. The Percentage of Recommendations for Low, Medium and High Recommendation ZIP Codes for Setting up a Restaurant.

The decision tree technique can be used to identify the impact of changes on results when one of the underlying attributes is changed. As the security and severity for a location changes constantly, this technique allows businesses to identify the factors that are more sensitive and less sensitive contributing to the security of the location. This kind of sensitivity is difficult to detect in another model.

4. FUTURE WORKS

Because there are many different algorithms and topics to explore in model creation and feature selection, it would be interesting to analyse the effects of these algorithms and verify if any of them could perform better on the data sets that we have.

Additionally, we can perform more robust analysis and develop a model with more accuracy if we have the ability to input additional information like housing rates, population demographics, accessibility to public transit, green space, school rating and population diversity. Due to time constraints, as well as the limit on resources and availability of public data, our analysis was restricted.

As the work is done on a local pseudo-distributed mode cluster and there is no significant speed up, Using a bigger cluster with more data on cloud computing service like Amazon Web Service maybe a good option, which will give an opportunity to use more big data technologies and use of lightning-fast speed of the cluster.

REFERENCES

[1] Nitish Gupta, Sameer Singh, Collectively Embedding Multi-Relational Data for Predicting User Preferences https://arxiv.org/pdf/1504.06165.pdf
[2] Using decision tree analysis to identify risk factors for relapse to Smoking, by Megan E. Piper, Wei-Yin Loh, Stevens S. Smith, Sandra J. Japuntich, Timothy B. Baker, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2908723/pdf/nihms176002.pdf
[3] Scott L. Minkoff, NYC 311: A Tract-Level Analysis of Citizen-Government Contacting in New York City, https://www.researchgate.net/publication/274708724_NYC_311_A_Tract-Level_Analysis_of_Citizen-Government_Contacting_in_New_York_City
[4] Apache Hadoop http://hadoop.apache.org/
[5] Apache Spark http://spark.apache.org/
[6] NYC 311 Calls Data https://data.cityofnewyork.us/dataset/311-Service-Requests-From-2011/fpz8-jqf4
[7] NYC Crime Reports Data https://data.cityofnewyork.us/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg
[8] NYC Restaurant Data https://www.yelp.com/dataset
[9] Introduction to Data Mining, Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Pearson Publishing
[10] Ubuntu Download resource https://www.ubuntu.com/download/desktop
[11] Hive https://hive.apache.org/
[12] Decision tree analysis for the risk averse organization. Hulett, D. T. Paper presented at PMI® Global Congress 2006—EMEA, Madrid, Spain. Newtown Square, PA: Project Management Institute. Also https://www.pmi.org/learning/library/decision-tree-analysis-expected-utility-8214