An Analysis of the Location Data of Hospitals in Metro Manila

Gabriel Ureta

https://github.com/gpureta/Coursera_Capstone.git

1. Introduction: Business Problem

When there is an opportunity to open a business, the existing infrastructure affects the type of costumers present. In the Philippines, government hospitals are more affordable than private hospitals but are less accessible due to long lines and schedules. Proximity and specialty aside, private hospitals cater to higher income patients by providing more comforts and amenities at a higher price.

Since the class (government or private) of the hospital affects the demographic, it would be worthwhile to learn how this affects the type of businesses and venues that exist at a walking distance from the hospital. Those visiting the hospital are also potential customers for businesses that are a walking distance from the hospital. Using the foursquare API, data about the venues surrounding government and private hospitals was compared to see if there was a meaningful difference. Finally, machine learning was applied to predict whether a hospital is of the class government or private based on the businesses and venues that surround it. The converse implication of this can be of use to a stakeholder when opening a business near a hospital, in that certain types of businesses are more likely to be present (if not more successful) depending if the hospital is public or private.

2. Description of Data Set

The hospitals used in this project are those located in the National Capital Region of the Philippines, which contains the cities and municipalities compromising Metropolitan Manila. This is to ensure that there is enough data from foursquare about the different venues surrounding the hospitals. A CSV file from the Philippine Department of Health website already contains the list of health facilities as well as their respective street address. This had to be filtered by choosing relevant facilities with types “Hospital” and “Infirmary”. **Geopy** was used to get the specific latitude and longitude for each hospital using the street address. The foursquare API was then used to gather information about the different venues surrounding each hospital.

The specific data gathered about the venues were the category and the number of venues. The idea is that Government Hospitals will attract a different set of businesses from Private Hospitals. The venues gathered was set at 500 meters, which is the maximum comfortable walking distance from a hospital. The CSV file from the department of health also contains data about each hospital's bed capacity, which was also used for an analysis on the effect of the size of the hospital on the businesses that surround it. Hospitals with no listed bed capacity were dropped from the data set for this analysis.

Folium was used to create an interactive map to visualize the data and to drop hospitals that were outside of the study area. Some of the other hospitals that were dropped during the course of the project were government mental health facilities, prison hospitals, and other facilities which are either outliers or irrelevant to the problem.

Table 1: Sample of the data set

	Name	Type	Class	Street	City	Beds	Latitude	Longitude
0	ALABANG MEDICAL CENTER	Hospital	Private	ALABANG-ZAPOTE ROAD	MUNTINLUPA	18.0	14.419099	121.043914
1	BERNARDINO GENERAL HOSPITAL II	Hospital	Private	BLOCK 1, LOT 2, NORTH OLYMPUS SUBDIVISION, ZAB...	QUEZON CITY	35.0	14.700589	121.034970
2	CHINESE GENERAL HOSPITAL AND MEDICAL CENTER	Hospital	Private	286 BLUMENTRITT STREET	SANTA CRUZ	592.0	14.626311	120.987791
3	DE OCAMPO MEMORIAL MEDICAL CENTER	Hospital	Private	2921 NAGTAHAN STREET SANTA MESA	MANILA	20.0	14.599832	120.999627
4	DR. FE DEL MUNDO MEDICAL CENTER	Hospital	Private	11 BANAWE STREET	QUEZON CITY	107.0	14.620604	121.009230

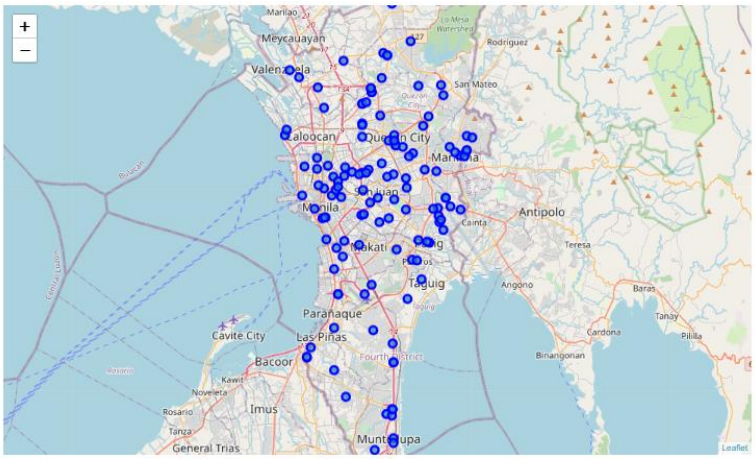

Figure 1: Interactive map of the data

3. Methodology

Previously, the gathering of the data for the venues that are a walking distance from each hospital was discussed. The data for the bed capacity of each hospital was also readily available. The next step was to perform some exploratory data analysis. The question to be answered in this step was:

- What are the differences between the type of businesses/venues that surround government and private hospitals?

In this step, the Pearson correlation factor was used to see if the size of the hospital (dictated by its bed capacity) affects the type and number of venues. This was done to see if the data for each hospital had to be scaled according to their bed capacity. An analysis of variance was also used to see if the difference of the type of venues for government and private hospitals were statistically significant.

The final step included machine learning classification systems. The question to answer in this step was:

- Can we predict if a hospital is government or private based on the venues that surround it?

Here, data was split into training sets and testing sets in order to verify the out-of-sample accuracy. The classifier algorithms used was ‘k-nearest neighbours’, ‘decision trees’, ‘support vector machine’ and ‘logistic regression’. Each classifier algorithm was optimized by choosing the best parameter (e.g. the best 'k' in K nearest neighbours). The performance of each algorithm was then compared.

4. Results

4.1. Exploratory Analysis

First, the effect of the size of the hospital on the surrounding venues needs to be taken into account. A larger hospital might cause a greater number of venues due to a higher potential customer traffic. The data used to represent the size was the hospital bed capacity. The number of restaurants and pharmacies located at a walking distance (500 meters) were then compared. These two venue categories were chosen because they are the most relevant to hospital data, and to reduce the number of API calls.

4.1.1 Beds vs Pharmacies

The Pearson Correlation Coefficient measures linear correlation between two sets of data. A value closer to +1 indicates a positive linear correlation and a value closer to −1 indicates negative linear correlation. A value near zero indicates zero correlation.

Table 2: Bed Capacity vs Number of Pharmacies

Correlation Coefficient	0.2741700569034865
P-value	0.2741700569034865

The correlation coefficient is close to zero which implies that there is no linear correlation between the bed capacity of the hospital and the number of pharmacies. This can be seen in the scatter plot (figure 2), which is not linear. The P-value however is greater than 0.1 which indicates that there is no certainty in the data.

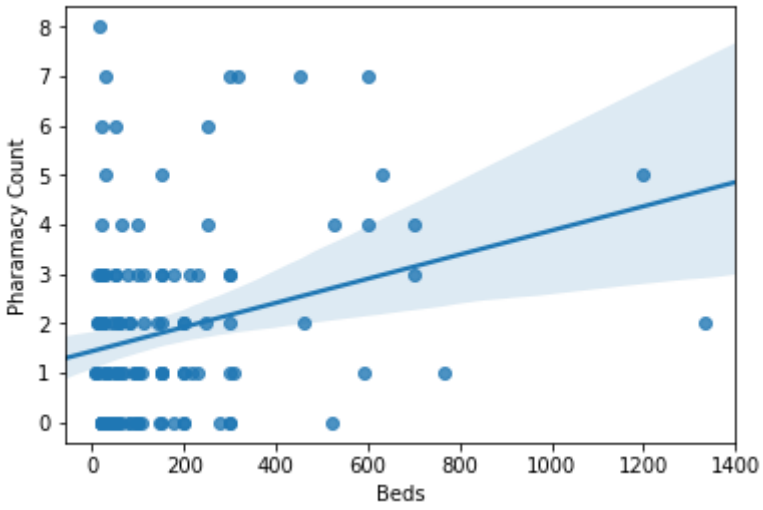

Figure 2: Scatter plot of bed capacity versus the number of pharmacies

4.1.2. Beds vs Restaurants

Table 3: Bed Capacity vs Number of Restaurants

Correlation Coefficient	0.2850344176504797
P-value	0.2850344176504797

The correlation coefficient is close to zero which implies that there is no linear correlation between the bed capacity of the hospital and the number of restaurants. This can be seen in figure 3 below, which is not linear. The P-value however is greater than 0.1 which indicates that there is no certainty in the data.

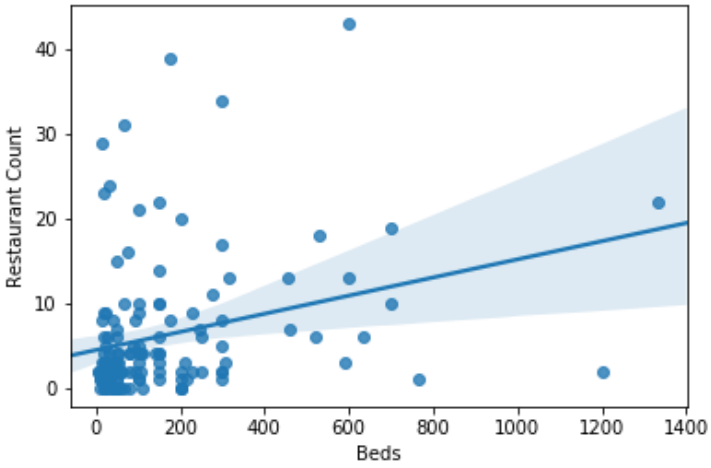

Figure 3: Scatter plot of bed capacity versus the number of restaurants

Since it looked like the bedspace capacity had no linear correlation on the number of venues, it was not included as a feature in the succeeding analyses. There are a lot more factors that affect the number and type of venues around the hospital. In fact, the hospital might not be a factor at all, but this was just done in order to perform exploratory analysis and to see if it is a viable feature to add in the next step. It could also just be that there is just not enough data (high P-value).

4.1.3. ANOVA

In this step, an analysis of variance (ANOVA) was performed on the data to compare the means of the Goverment and Private hospitals for each category. The venues for each hospital were gathered using the foursquare API and were then summed up by category. This was processed using one hot encoding. After applying ANOVA using f_oneway from scipy.stats, the categories that were significant were found to be:

Table 4: P-values of significant categories

Chocolate Shop	0.036153
Food Court	0.003575
Fountain	0.036153
Indian Restaurant	0.036153
Museum	0.047205
Optical Shop	0.009674
Scenic Lookout	0.036153

A P-value with less than 0.05 suggests that the class (government/private) of the hospital affects the number of the specific venue category. It seems like the class of the hospital affects the number of chocolate shops, food courts, fountains, Indian restaurants, museums, optical shops, and scenic lookouts. The lowest P-value is that of Food Courts with 0.003575. Correlation does not equal causation, and all of these values can just be happenstance. There is not much to learn here, except that there is moderate certainty that there are less food courts in private hospitals than in government hospitals.

4.2. Classification

In this section, machine learning algorithms are used to see if they can predict the class of the hospitals are used based on the venues that exist around it. The data was split into 80% training data and 20% testing data to check the out-of-sample accuracy. The data was also normalized so that outliers won't affect the data and different scales for each feature would be addressed (this step was optional because the data are all in the same units, but it is good practice to normalize). The classification models were fitted using the training data, and the accuracy was tested using the testing data. After the best model was achieved, the model was then trained using the whole data and tested against the testing data.

4.2.1. K Nearest Neighbors (KNN)

KNN is an algorithm that uses K closest values for prediction. In order to have the best accuracy, different values for K were tried.

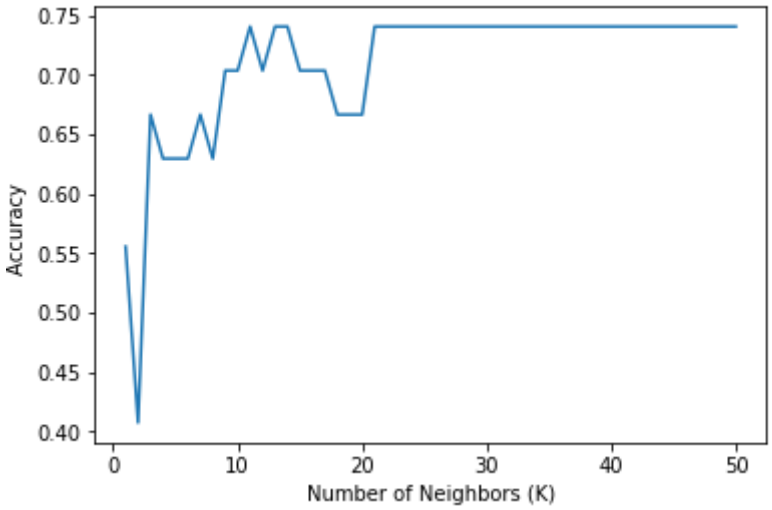

Figure 4: Model accuracy for different K values

The highest accuracy was 0.74 at k = 11 and the same for K values of 22 and onwards. Inspecting the plot, the accuracy elbows at around K = 22. This is usually a better K to use. Testing the accuracy with k = 22 and using the whole data set (instead of just the training data) results with an accuracy of 0.78. However the recall (True positive rate) for Government hospitals is very low at 0.14.

Table 5: Confusion Matrix with K=22

	precision	recall	f1-score	support
Government	0.50	0.14	0.22	7
Private	0.76	0.95	0.84	20
micro avg	0.74	0.74	0.74	27
macro avg	0.63	0.55	0.53	27
weighted avg	0.69	0.74	0.68	27

4.2.2. Decision Tree

Decision trees classify labels by learning simple decision rules inferred from the data features. The deeper the tree, the more complex the decision rules to fit the model. The model was tested at different depths.

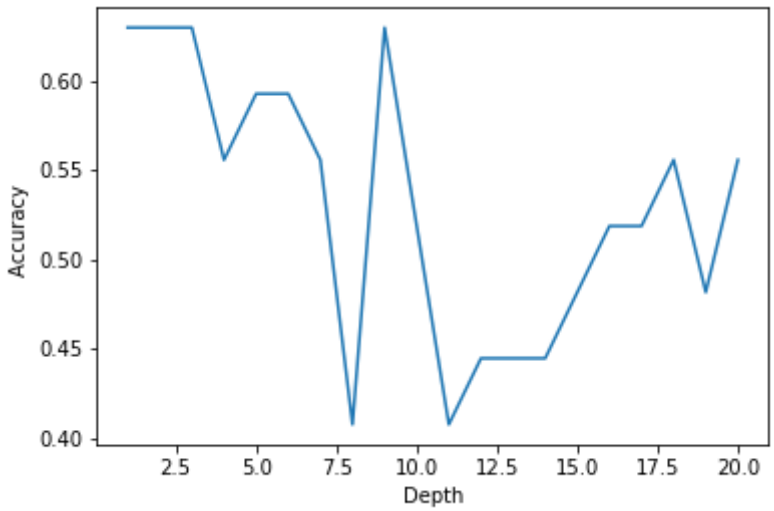

Figure 5: Model accuracy for different depths

The highest accuracy was 0.63 at depth = 1 and depth = 9. The depth was set to 9 since a depth of 1 might be too simplistic and be underfitting. Using the whole X data set, the accuracy is around 0.78 with good recall scores compared to KNN.

Table 5: Confusion Matrix with depth = 9

	precision	recall	f1-score	support
Government	0.56	0.71	0.63	7
Private	0.89	0.80	0.84	20
micro avg	0.78	0.78	0.78	27
macro avg	0.72	0.76	0.73	27
weighted avg	0.80	0.78	0.79	27

4.2.3. Support Vector Machine (SVM)

SVM performs classification by finding the hyperplane that separates the data. Mapping data into a higher order space can be done using different kernelling models. In this step, different models were tried and the models with the highest accuracy was found to be 'poly' (polynomial) and 'rbf' (radial basis function), both at around 0.74. Applying both models and fitting the whole X data set, they both have an accuracy of 0.74.

4.2.4. Logistic Regression

Logistic regression classification uses a logistic function to model a binary dependent variable. It can use different algorithms to solve the optimization problem. The best solver was 'newton-cg' with an accuracy of 0.704. The solution did not converge for the 'sag' and 'saga' solvers. Using 'newton-cg' on the whole X dataset to train the model, the accuracy on the test data set was around 0.74. However, the log-loss value (lower is better) was only at 0.5.

5. Discussion

For the exploratory analysis, it looked like the bedspace capacity had no linear correlation on the number of businesses surrounding the hospital. Although it is logical to think that there should be more venues for a larger hospital, there are a number of reasons of why this might be:

- The number of patients visiting the hospital is not proportional to the bed capacity. Outpatient services that do not require beds can be a factor.
- The location demographics can affect the type of businesses present. For example, a highly urban city like Makati can have a high traffic of potential customers, not necessarily from patients. A hospital located in a major highway or near a large shopping mall will also have higher number of venues surrounding it, regardless of its bed capacity. In this case, the hospital isnt even a factor at all.
- There simply isn't enough data to test the hypothesis, either from the number of venues or the number of hospitals. With a large enough data set, a higher certainty can be achieved.

Ironically, this made the subsequent analyses easier, since the venues data no longer had to be scaled down based on the hospital capacity.

The same problems are true for the analysis of variance. Between government hospitals and private hospitals, the only difference with a moderate certainty was the number of food courts. It can be assumed that since government hospitals cater more towards lower costs, they have more food courts than private hospitals which would be more inclined towards having restaurants.

For the machine learning classification systems, all the models performed well. Each model had a test accuracy of about 70%, with K nearest neighbors having the best test accuracy at around 0.74. Training using the whole data set, the accuracy increased to 0.78. However, even after training the model using the whole data, the recall (True Positive Rate) for government hospitals was only 0.14. The Decision Tree had the lowest test accuracy at only 0.63, but after training using the whole data set, the accuracy reached 0.78 and did not suffer low recall like KNN.

6. Insights and Conclusion

Answering the first questions laid out for the project, the only discernable difference between the type of businesses/venues that surround government and private hospitals was the number of food court venues. The size of the hospital was not a factor in the number of venues that surround it. This shows how actual data can be different from the expected results, and the importance of exploratory analysis.

Answering the second question, predicting if a hospital is government or private based on the venues that surround it is possible. The best model from those used was the Decision Trees Classifier, although each model performed relatively well.