Using Machine Learning to predict frequency of COVID-19 cases in Toronto

Mirko Simunovic

August 10, 2021

1. Introduction

The current pandemic of the COVID-19 virus has impacted all aspects of daily life, businesses, international relations, travels and families. As a consequence, governments and health national agencies need all the information available to be able to predict a potential new focus of infection. This is a hard task because information about COVID-19 spreads is not readily available to authorities and depend mostly on lab tests which can take days to confirm when a person is infected. For this reason, governments and health agencies are sometimes too late behind in taking the correct sanitary measures in a given urban area. Therefore, being able to predict new focus of infection would prove to be very valuable in the fight against COVID-19. We will create a machine learning (ML) model that uses the location data from previous focuses of infection in the city of Toronto and determine characteristics that can be used to identify potential new neighborhoods where infection can spread in similar ways. These characteristics can include, for example, commercial activity in the neighborhood, outdoor sport venues, parks and shops, as well as demographic information (population, median income, etc). The model will learn which neighborhoods are likely to spread the virus and become focuses of infection given the types of activities that are commonly held by their communities.

2. Data

We used a database of COVID-19 infections in Toronto. This data set contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020. This includes cases that are sporadic (occurring in the community) and outbreak-associated. The data are extracted from the provincial communicable disease reporting system (iPHIS) and Toronto's custom COVID-19 case management system (CORES) and combined for reporting. The complete database can be obtained from www.kaggle.com in csv format.

The location venue data was obtained from the Foursquare API. With the API, we can obtain detailed and complete lists of venues that are found in each neighborhood, within a given distance radius. These data provide the ML model with information about which activities commonly take place in the neighborhood.

Finally, we used a database of demographic information about all the neighborhoods in the city of Toronto. These neighborhood profiles were developed by the Community Data Program (https://www.toronto.ca/city-government/data-research-maps/) to help government and community agencies with their local planning, by providing socio-economic data at a meaningful geographic area. From the neighborhood profiles, we have used information about total population and income distribution in each neighborhood.

The final data sample consists of 140 instances (neighborhoods), each with many tens of columns, as explained later.

3. Methodology

3.1. Data Preparation

There was several problems with the COVID-19 cases table. One important issue is to list the Toronto neighborhood names in the same exact format for all tables. This was done manually when necessary. Some other relevant data preparation steps are to reject all the COVID-19 cases of suspected infection and only keep the confirmed cases. Finally, we have to drop all rows of null data in the COVID-19 cases table since we cannot fill the missing data.

For the venue location data we have used the geopy package to obtain the location of all Toronto neighborhoods. Their location is shown in Figure 1 below.

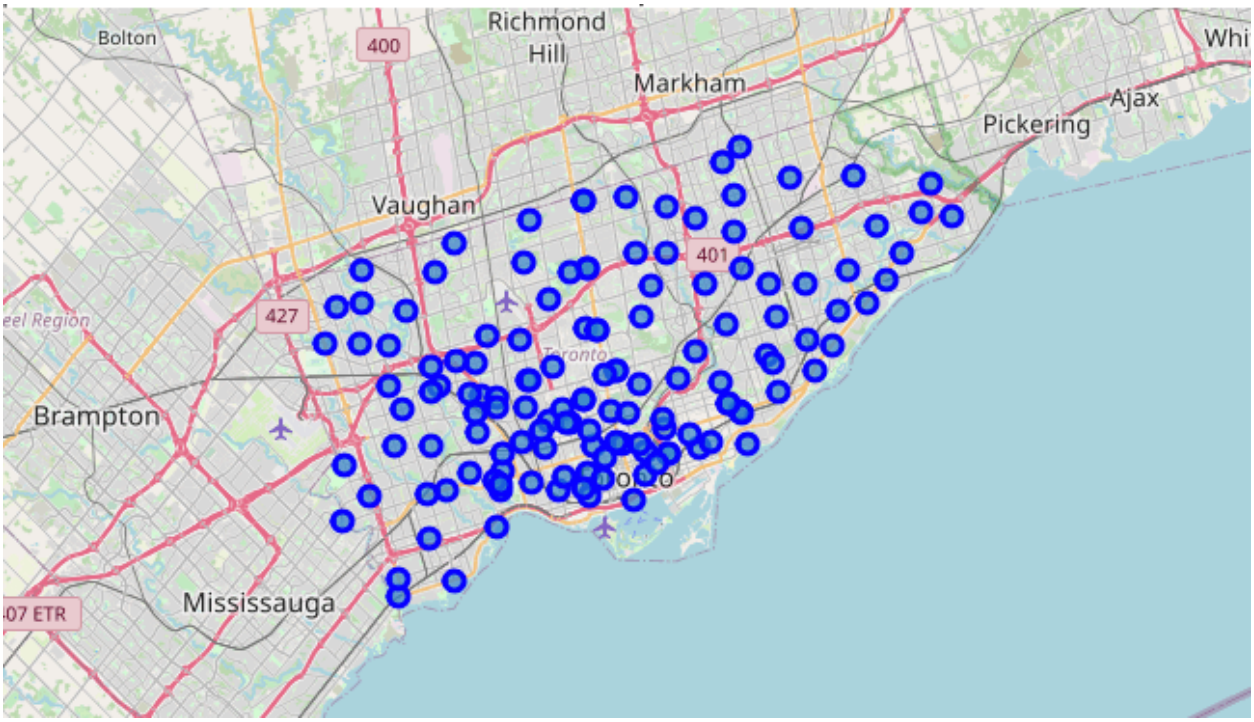


Figure 1. Location of all neighborhoods in the city of Toronto. The locations are obtained from the geopy package.

We used the location data to extract all the venues in a radius of 1000 meters. The data comes from the Foursquare API. This way we obtain more than 6,000 venues across all neighborhoods. The 5 more popular types of venues are: Coffee shop, Cafe, Park, Pizza Place and Italian Restaurant.

For the demographic data we need to rename some of the neighborhood names to match the names in the infection table data. This data table contains, among others, the population in each neighborhood and the amount of people that have a yearly income inside a given bracket. The brackets go from $5,000 until $100,000 and over. For each bracket we have the total number of people in each neighborhood. We had to make a function that calculates the cumulative number per bracket, and then we use that function to estimate the 20, 50 and 80-percentile income levels.

3.2. Data Exploration

Now we want to explore the data and look for interesting insights about the way that COVID-19 infections are occurring in Toronto.

- Total Cases vs Number of Venues

We show in Figure 2 the relation between total cases and total venues in each neighborhood. At first inspection, the result is very interesting. Contrary to our expectation, the neighborhoods with the more infections actually have little number of venues, and viceversa.

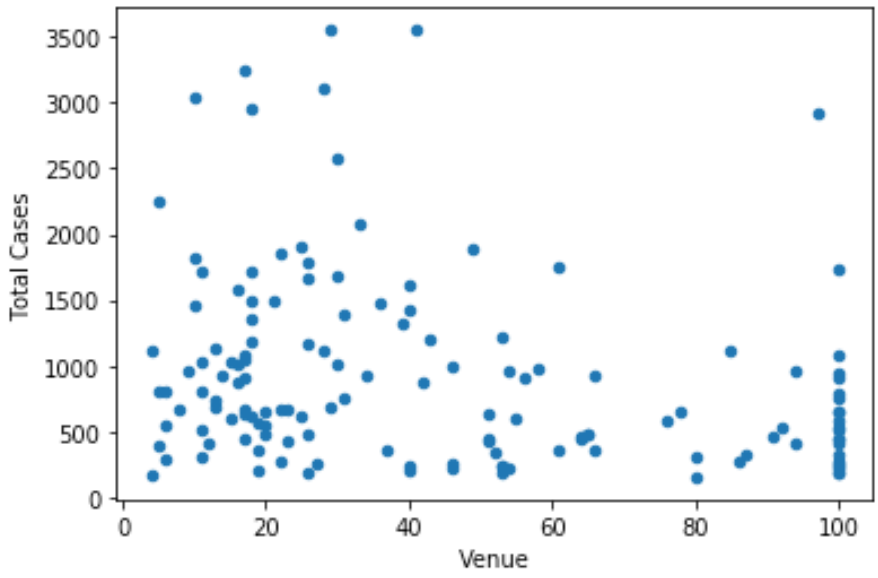


Figure 2. Total Venues versus Total Cases per neighborhood.

We have also looked at the correlation between Total Cases and the number of specific types of venues.

Venue Type	Pearson correlation
Café	-0.320584
Total Venues	-0.303757
Bakery	-0.299674
Ice Cream Shop	-0.299095
Burger Joint	-0.297114
Sushi Restaurant	-0.294087
Italian Restaurant	-0.289383
Bar	-0.249326
Thai Restaurant	-0.247097
Pub	-0.240891
Breakfast Spot	-0.240386
Mexican Restaurant	-0.234243

The 5 more correlated types of venues are cafe, bakery, ice cream shop, burger joint and sushi restaurant. They are all anti-correlated, so this shows again that more infections are occurring in neighborhoods with little commercial activity. We show in Figure 3 a box plot of the distribution of Total Cases, when grouping neighborhoods by total number of Bakeries and Cafes.

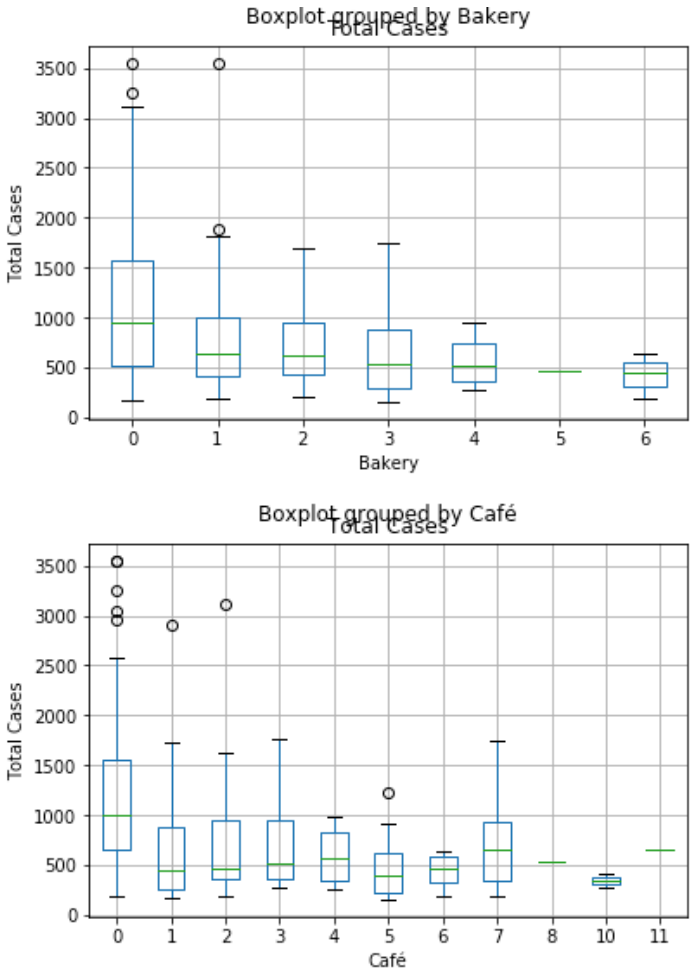


Figure 3. Top: Total cases boxplot distribution for neighborhoods grouped as a function of number of bakeries. Bottom: the same but for number of cafes.

- Total Cases per thousand habitants

Using the demographic table, we can normalize the number of cases in each neighborhood to the number of infection per thousand habitants. This gives us a statistic that is more meaningful and can be more robust for ML modeling. Now let's see the same figure using the normalized infection rate.

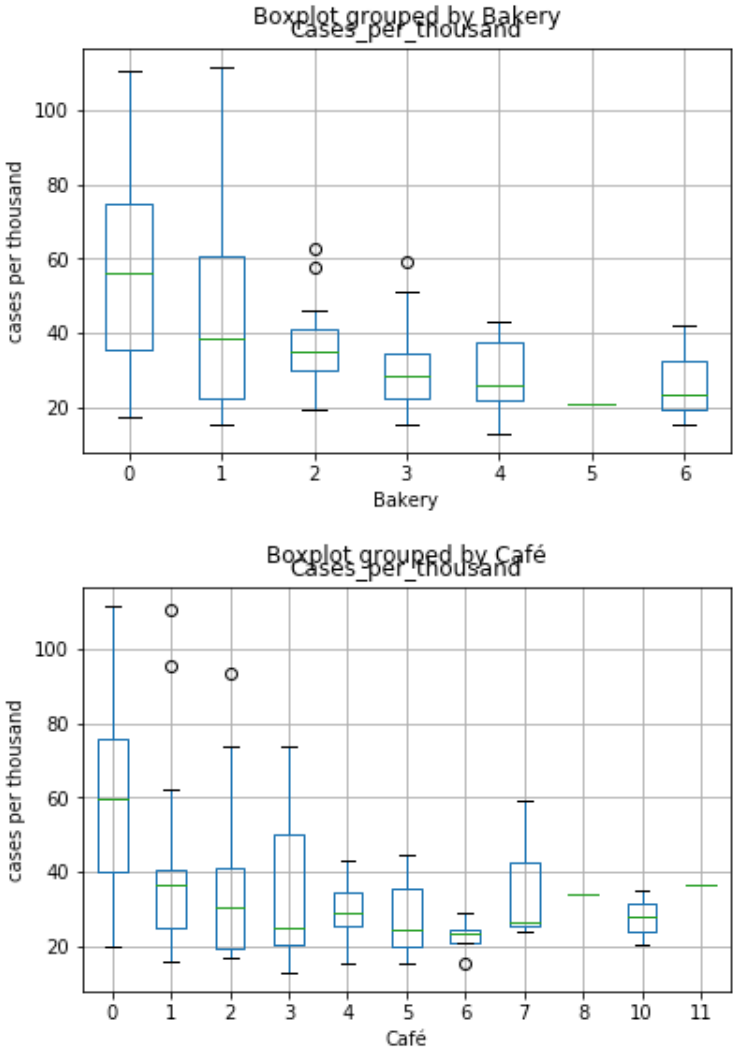


Figure 4. Same as Figure 3 but for normalized infection rates.

- Income Statistics versus Infection Rates

We now include the financial income information into the data exploration. The first test is to calculate again the Pearson correlation coefficients of every column against the normalized infection rate (cases per thousand).

Table column	Pearson correlation
household_fourthq_income	-0.588745
Total Venues	-0.501046
Café	-0.429447
Sushi Restaurant	-0.424669
Bakery	-0.411952
Italian Restaurant	-0.364729
Park	-0.342780
Pub	-0.338928
household_med_income	-0.336033
Thai Restaurant	-0.333169
Coffee Shop	-0.332425
Bar	-0.324370
Restaurant	-0.310035

We see that the correlations coefficients are now much more statistically significant. In fact, the fourth quartile income level is strongly anti-correlated to the infection rates. Its correlation coefficient is stronger than for any venue type. This gives more evidence to the observation that infection rates seem to be strongly dependent on economic activity and living standards. The initial hypothesis is that infections are numerous in zones of low income due to possibly the crowding in households, higher density from apartment buildings, and likely more exposition to public transportation.

We show in Figure 5 the fourth quintile income level per household of each neighborhood as a function of the infection rate. The figure confirms very clearly that neighborhood with high income are much more likely to have low infection rates. Conversely, the highest infection rates occur only in neighborhoods with the lowest income.

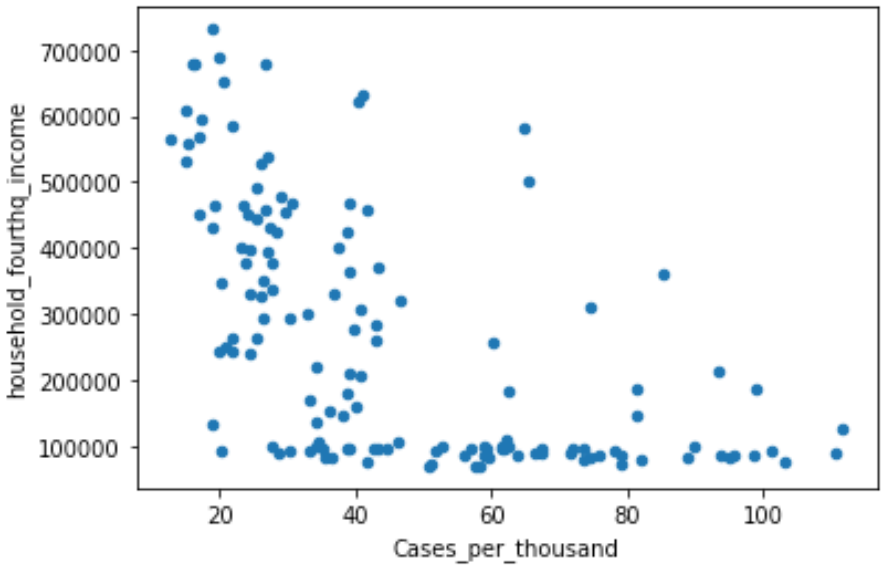


Figure 5. Scatter plot of the the fourth quintile income level per household of each neighborhood as a function of the infection rate.

- Normalizing all columns to number per thousand habitants

We observe that the number of venues are giving us important information about the types of commercial and economic activity in the neighborhoods, but this quantity must be normalized because logically some neighborhoods are much more populated than others, so it makes sense to see the number of venues per

thousand habitants. Therefore, we create new table features by normalizing every column about number of venue types, as well as the columns about income statistics (the income at 20, 50 and 80-percentile level.).

3.2. Building Machine Learning Models

For the machine learning models we need to use Regression Models. This kind of models are chosen because we will try to predict a unique numeric value: the infection rate at each neighborhood. The features we will use are all of the Venue information columns which contain the number of type-venue per neighborhood (and its population normalized duplicate as well), and also the income statistics: the median income, the first quintile and fourth quintile income levels (and also their population normalized duplicated columns). Another possibility would have been to classify the infection rates into a, for example, three class hierarchy (low, medium, and high infection). Then we could have used a multi-class classification model. However, we prefer a regression model because we want to be able to predict the exact infection rate, not a broader category, even if the accuracy might be higher for a three class model.

First, we split the data into a train set and a test set. The test set is selected randomly from 30% of the full data, and it will be used to test our best final model. All the features are then scaled with a Standard Scaler (i.e. have a zero mean and sigma equal to 1). For our error function, we choose the root mean squared error (RMSE) in all models. All the models are built with the sklearn package in python.

We begin with (1) a simple Linear Regression model, and continue adding more complexity to the models, with (2) 3rd-degree Polynomial Regression, (3) Linear Support Vector Machine (SVM), (4) Decision Tree Regression and a (5) Random Forest Regressor model.

For all 5 models, we use a 10-fold cross-validation during the training, so that we can obtain a robust measure of the RMSE of each model by taking the average of the 10 folds. This is very important because our data set is very small (only 140 instances for the full data, and 98 training instances) so the risk of overfitting the training data is significant. For this, having cross validation can help us to make sure that the models can be correctly generalized. The results of the RMSE for each model are shown in the next section.

3. Results

The results are shown in Table 1. For each of the 5 regression models, we show the average RMSE from the 10 folds during cross-validation and its standard deviation (sigma).

	Linear Regression	Polynomial Regression	Linear SVM	Decision Tree Regressor	Random Forest Regressor
avg. RMSE	33.5	49.2	**20.4**	23.3	**17.5**
sigma	11.7	47.8	**4.9**	6.5	**5.3**

Table 1. RMSE statistics from 10-fold cross-validation in each regressor model.

As it can be seen, the best model is the Random Forest Regressor and it performs significantly better than all other models. The second best is the Linear SVM model.

3.1. Optimization of hyper-parameters of Random Forest Regressor

Now that we have found that a Random Forest Regressor is the best performing model, we can proceed to tune the hyper-parameters and see if we can get a better model. For this we use the GridSearchCV module in sklearn. We use a parameter grid which varies the hyperparameter "n_estimators" (number of Decision Trees used in the model) and "max_features" (fraction of features used from the train data). The grid values are shown in Table 2.

Hyperparameter	Values in Grid								
n_estimator	5	10	15	20	25	30	50	100	150
max_features	0.2	0.4	0.6	0.8	1.0				

Table 2. Hyperparameter grid for the Random Forest Regressor Model

The grid size is 9x5=45 different models, and each model is trained 10 times (10-fold cross validation). The grid search gives us the best RMSE for the set of

hyperparameters n_estimator=25 and max_features=1.0. We obtain for this model an average RMSE=17.33. Note that this is better than the average RMSE that we obtained for the initial Random Forest model with default hyperparameters.

Now we can take this model and use it to predict the infection rates in the test data set. This is a crucial step because we note that the test data set has not been used in the model training and therefore it is a completely new set of data. If the model performs much worse than for the train data, that is a sign that we could be overfitting the model. The comparison between predicted infection rate and real infection rates is shown in Figure 6.

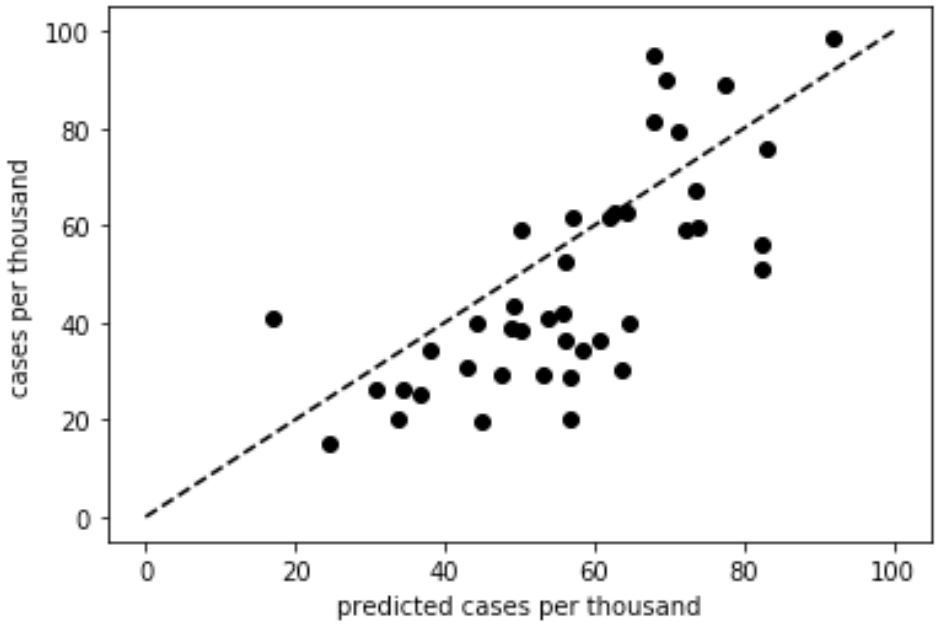


Figure 6. Comparison of predicted versus real infection rates from the Random Forest Regressor model.

The model does a good job and makes reasonable predictions for the entire range of values. We find that the RMSE of the test data set is RMSE=17.35. This is a very good value since it is in great agreement with the average RMSE from the cross validation process. Therefore the model is not showing signs of overfitting and it is generalizing appropriately to new data.

We can also look at the predicted infection rates in all 140 neighborhoods. This is of course not a measure of performance because was trained on that same

data but nevertheless it is useful to see how well the model can fit the data. This is shown in Figure 7.

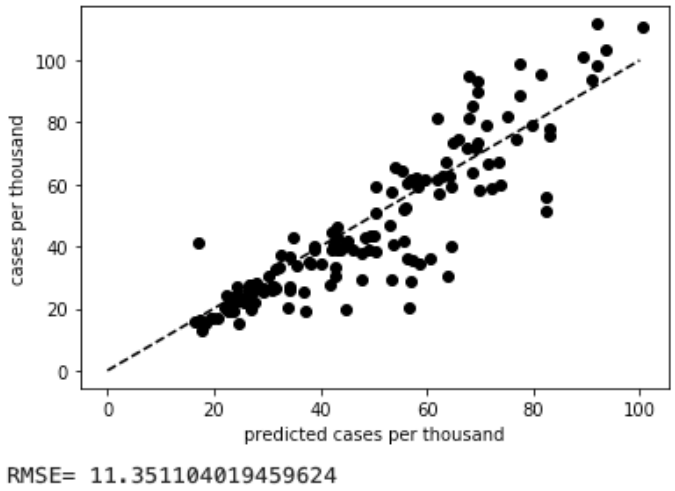
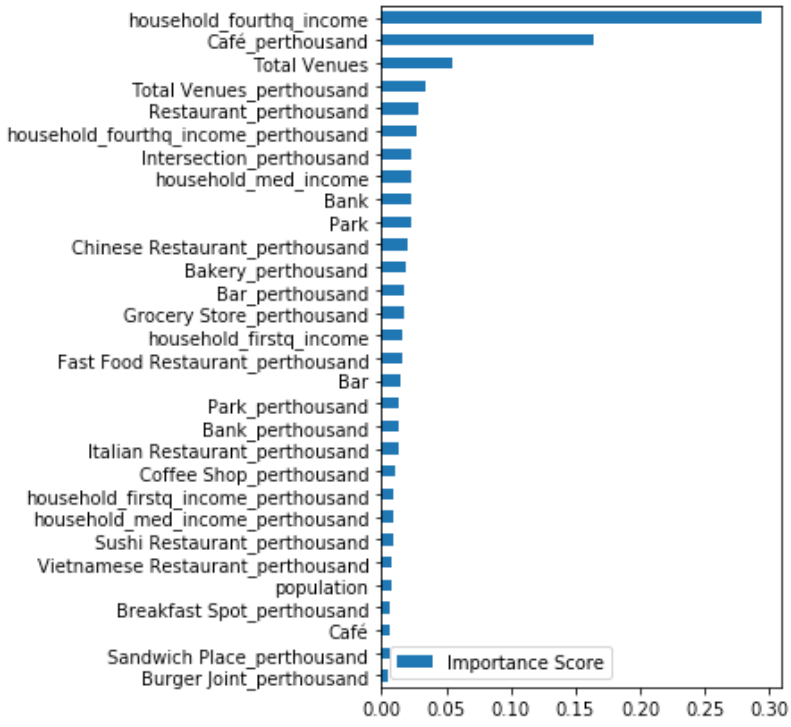


Figure 7. Same as Figure 6 but for the entire neighborhood data set.

As expected, the fit looks much better than Figure 6. The model is therefore able to accurately fit most of the data. We show now the relative importance of the top 30 features in the Random Forest Regressor model.



4. Discussion

The relative importance of the model features shows clearly that the most important information is the household fourth quintile income level. The second most important is the number of Cafe per thousand habitants. These two features reveal big insights about the cause of the observed infection rate in each neighborhood. The model suggests that infection rates depend most strongly on the living standards of the neighborhood. Areas of high income and high commercial activity have low infection rates, whereas areas of poverty and higher density have high infection rates, likely because of the worse living conditions and crowding in small housing units.

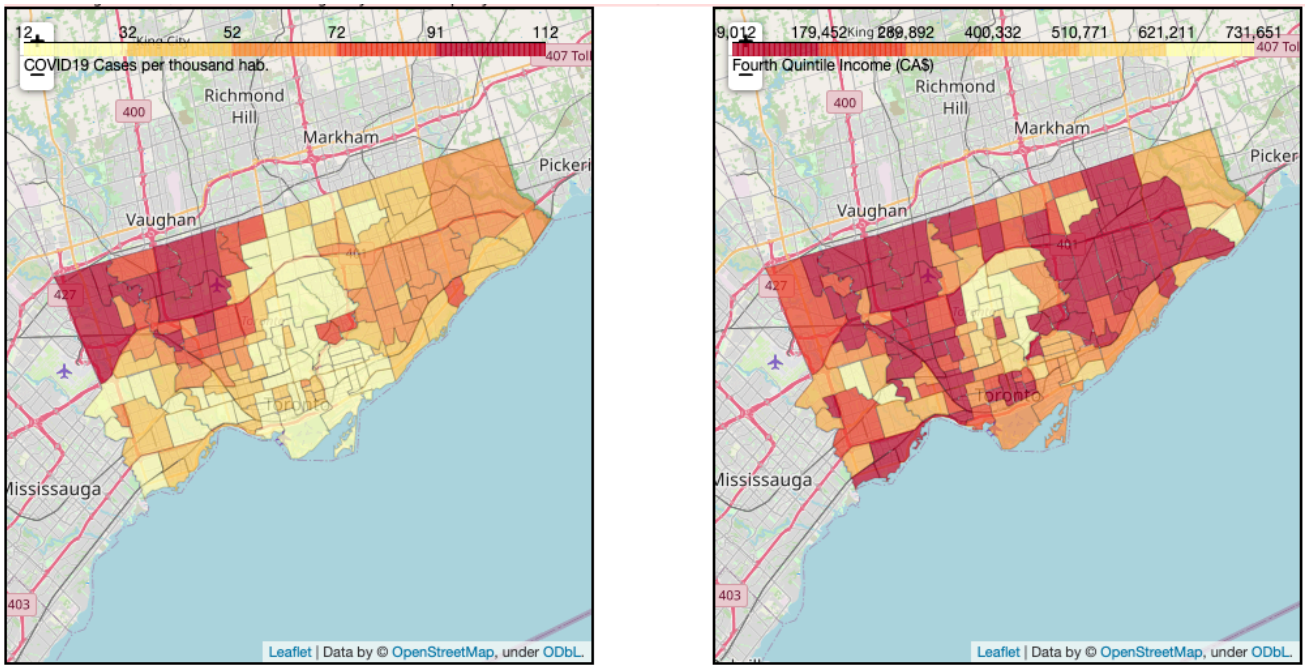


Figure 8. Left: Map of Infection rates in Toronto neighborhoods. Right: 80-percentile income level in each neighborhood.

We show in Figure 8 the map of the infection rates and the map of the 80-percentile (fourth quintile) income level. We can see some interesting patterns, where high income neighborhoods are clustered in the center of the city and extend toward the south and into the coastal area. A similar pattern is seen for the infection rates, where the East and West periphery clearly show higher infection

rates and the center and all the low COVID-19 infection areas are clustered in the center and south coastal areas. One notorious observation is the very low income at the North West neighborhoods, and the clustering of extremely high infection rates in the same area.

The model is therefore suggesting that COVID-19 infection rates will spike in neighborhoods of very low income, and thus the government authorities should focus their efforts in such areas.

5. Conclusion

We have used Machine Learning techniques to create a model of the COVID-19 infection rates of neighborhoods in Toronto. The data we have used is number and types of venues in each neighborhood as well as the 20-,50- and 80-percentile income levels and population in each neighborhood. The best model is a Random Forest Regressor, which is able to correctly predict the infection rate in neighborhoods with a RMSE of about 17 cases per 1,000 habitants. The highest COVID-19 infection rates can be about >100 cases per 1,000 habitants, so this is only a ~17% error in the predicted value for high infection areas. We have shown that Machine Learning can give authorities a number of great and valuable insights about the behaviors and activities that are most related to COVID-19 infections. Such statistical models should be at the base of any meaningful strategy of new health policies that aim to control the spread of the COVID-19 virus.