Testing the lifestyle difference with machine learning on the 2 sides of the ‘Latte Line’ of Sydney

Coursera IBM Data Science Specialization Course 9 Capstone Project

Author: Szabolcs Szombath

Date: 09/02/2020

Contents

1. Problem statement

1.1. What is the Latte Line?

Major cities around the world demonstrate a diversity in neighbourhoods in terms of lifestyle, income, social status and ethnic background of people, low- and high-density areas, and so on. In developed countries neighbourhoods are usually mixed based on the above criteria: ideally there are no ghettoes and no exclusive upper-class areas, however, the difference between districts can be quite apparent. In case of Sydney people are talking about a 'latte line' which divides the city to 2 parts, the prosperous North-East and the less attractive South-West. This division is well supported by statistical data: e.g. property prices, university admission rate of high school students, number of white-collar jobs tell the same story of geographical division. However, it is important to note that the population living South of the Latte Line is far not a homogenous mass and on average still has a high standard of living. This include many blue-collar workers as well: it is not uncommon for experienced tradies to have an income in the ballpark of $1000 a day.

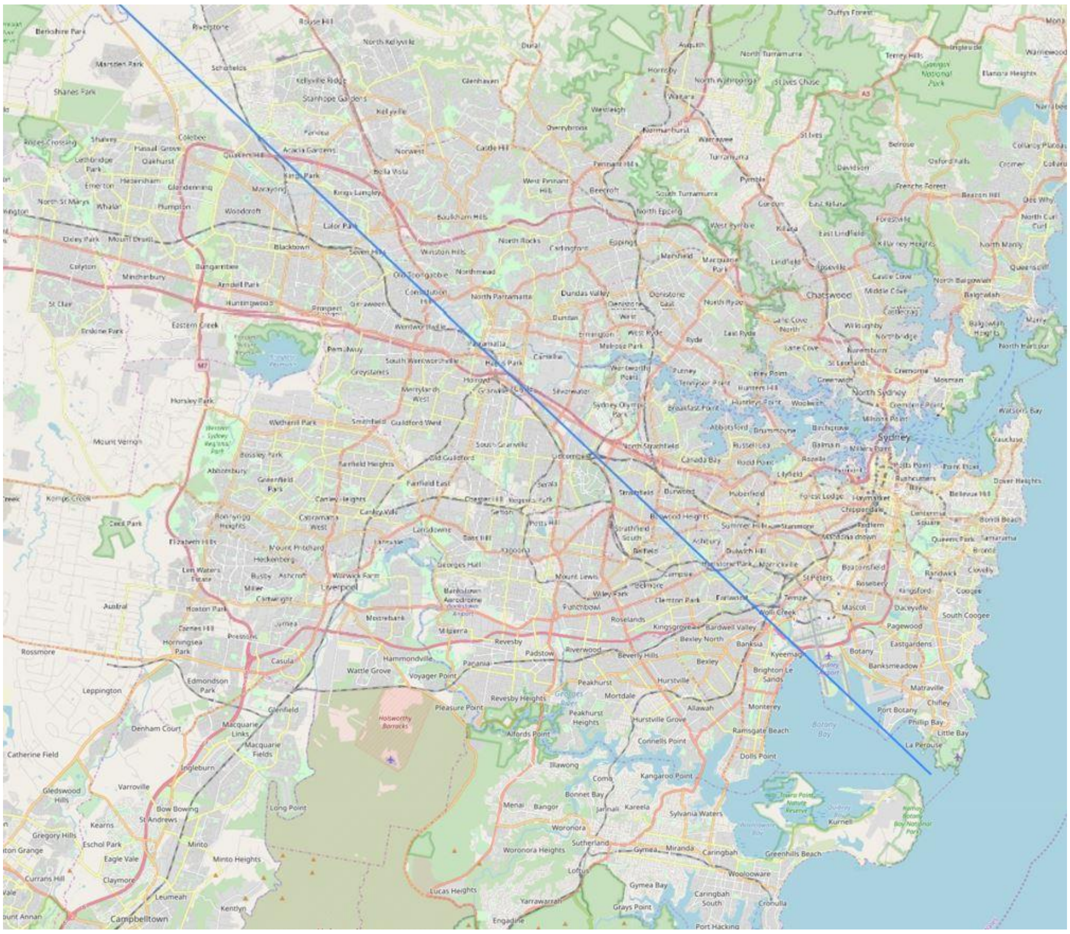


Figure 1: The Latte Line divides Sydney

1.1.1. Why is it called latte line?

The term was originally used in Tasmania to describe the difference between the posh North and the 'bogan' South. Etymologically latte is short for café latte which reflects visually the physical appearance of this type of coffee product having a distinguished fine white foam on the top and the dark liquid underneath; and also reflects the coffee culture of Australia with certified baristas brewing latte and flat white at a take away price of $4-5 for their demanding clientele. The term 'latte line' does not necessarily denote a difference in wealth, rather a difference in lifestyle (nota bene, lifestyle is often used as a euphemistic synonym for wealth, but in this case we focus on whether people on the 2 sides of the line follow a different style of life deliberately).

The question is, whether using Foursquare venue data, can we verify or debunk the 'latte line' as a division of lifestyles between 2 parts of the city with the help of data science?

2. Use of data to evaluate the 'latte line'

In order to use machine learning algorithms with regard to this question, the following data sets are needed:

- List of neighbourhoods
- Area and population size of neighbourhoods
- Geocode of neighbourhoods
- Foursquare venue list of the given areas
- Information on which neighbourhoods are above/below the line

The algorithm we are looking for should be able to classify neighbourhoods (North/South of the line) based on venues, thus we can handle this as a classification problem.

2.1. Questions to decide

2.1.1. Granularity of neighbourhoods

There is a wide choice of statistical area aggregates in the census package:

- mesh block
- statistical area 1-4
- suburb
- local governmental area

Out of these, mesh blocks are far too small for Foursquare venue lists, while statistical areas are difficult to handle: they are abstractions of the statistical bureau thus handling their geospatial data and verifying our results against the map would have additional complexity.

Suburbs and local governmental areas are easy to understand and find on the map, therefore these 2 are the convenient choices. Due to the low number of local governmental areas (thus the lack of data for proper classification) the only viable choice is the level of suburbs.

2.1.2. Matching neighbourhoods with Foursquare locations

The Foursquare API can return venues in a given radius of a given coordinate. A neighbourhood, on the other hand, normally have an irregular polygon shape. With the given set of information, it is not possible to perfectly match Foursquare data and suburbs, however, this is not necessarily a problem, since neighbourhood boundaries rarely influence people against visiting a venue, therefore the venue's potential impact on lifestyle is still present. Nevertheless, Foursquare circles and neighbourhoods should match as much as our data supports this.

2.1.3. Using venues as a mean of comparison

Venues can be matched against each other based on venue type. The significance of a venue with a given type on the assumed lifestyle of a neighbourhood is more difficult to grasp. Neighbourhoods are highly different in terms of population and area size, therefore, the sheer number of venues may not be the best indicator. Compering the number of venues as a ratio of *population* and *number of venues* is a convenient way to solve this problem.

2.1.4. Location of the Latte Line

For the sake of simplicity the line will be defined based on a single source, the geographical coordinates are taken from a report of the Australian national news agency, abc.com.au: https://www.abc.net.au/news/2019-12-17/sydneys-latte-line-divides-job-and-housing-opportunities/11803706. Even though different mediums may post somewhat different interpretations of the geographical position of the line, due to the arbitrary, theoretical quality of the location there is no value in further examination.

Note: a map is a 2D projection of a 3D planet, thus using latitude/longitude coordinates and linear regression is methodologically not entirely correct. However, considering the arbitrariness of the latte line, this error does not impact the success of labelling.

2.1.5. Success criteria

The goal of the analysis is building a model to predict if a neighbourhood based on the lifestyle represented by venues belongs to the North or South of the latte line.

Setting the accuracy threshold is arbitrary: while we cannot expect Foursquare data to make a definite distinction between latte/non-latte regions, if it is closer to 50% than 100% we may question whether venues really reflect a difference of lifestyle.

If we can achieve at least 75% accuracy we may decide that the model works.

2.2. Data sources

The following data sets will be used:

1. A list of neighbourhoods of Sydney
2. Population size and area size of neighbourhoods
3. Foursquare API with venues

For practical reasons the list of neighbourhoods will be taken from the Australian Bureau of Statistics, since we can get population size (number of residents) at the same granularity. Neighbourhoods are going to be assigned with their latitude/longitude coordinates with the help of the geopy API.

Neighbourhoods will be labelled on which side they are of the latte line based on their coordinates.

2.3. Normalising data

Data sets from the statistical bureau is joined to provide a single data frame with:

- Suburb name
- Population size
- Area size

The list contains all suburbs of New South Wales therefore filtered with the help of the Sydney suburb list.

2.3.1. Assign latitude and longitude columns to the data set

Before fetching Foursquare data, geocoding must be done, thus latitude and longitude is assigned, using the geopy API, the records will have:

- Suburb name
- Population size
- Area size
- Latitude
- Longitude

2.3.2. Radius

As with Foursquare we have to use radius while suburbs tend to have polygon shape, it is important to check the size and position of our 'technical' suburbs. Radius is calculated as if the total area of a suburb were covered by a circle ($r^2\pi$).

- Suburb name
- Population size
- Area size
- Latitude
- Longitude
- Radius

2.3.3. Assigning latte-line label to data set

To enable machine learning a major piece of information need to be added, whether a suburb us above or below the latte line. This is calculated with linear geometry, and the data set will have the following values:

- Suburb name
- Population size
- Area size
- Latitude
- Longitude
- Radius
- Latte

2.3.4. Check initial data set on map

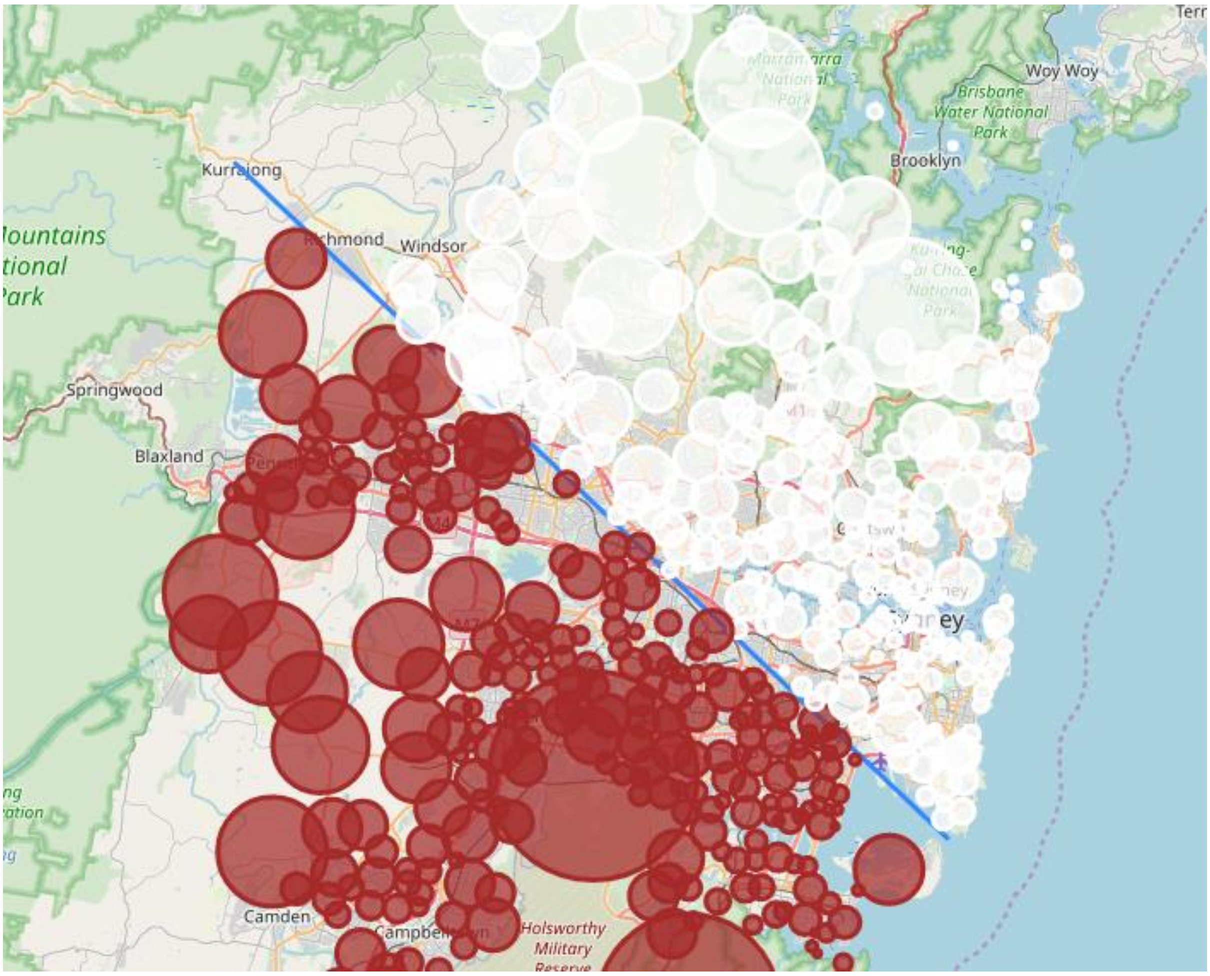


Figure 2: Geocoded suburbs with radius on both sides of the line

An important discovery is that based on the map, Holsworthy and Ku-Ring-Gai Chase must be removed to minimise autocorrelation.

2.3.5. Addressing data scarcity

Working with venue types occurring in low numbers is not useful for this exercise since those do not help comparisons. Venue types with less than 5 instances are removed.

Suburbs with extreme low number of venues cannot add value to the exercise since there is no comparable data across the regions. Suburbs with less than 10 venues are removed.

These 2 steps half the number of venue types and suburbs eligible for further analysis.

2.4. Decision on algorithms

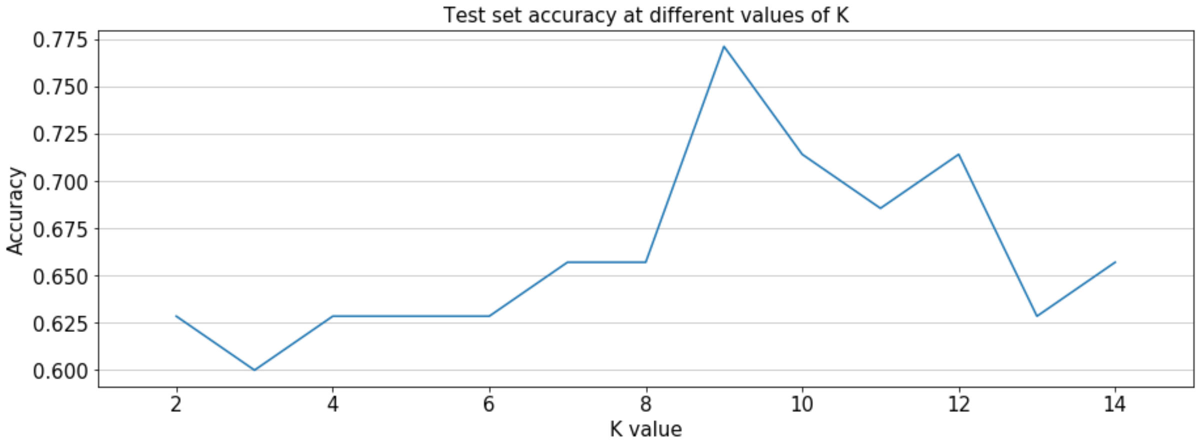
During the previous courses the following 4 algorithms were used for classification problems:

- K Neighbours
- Logistic regression
- SVM Model
- Decision Tree

Since the data set is relatively small, all 4 of them can be safely executed, including optimisation of processing parameters.

2.4.1. K Neighbours Classifier

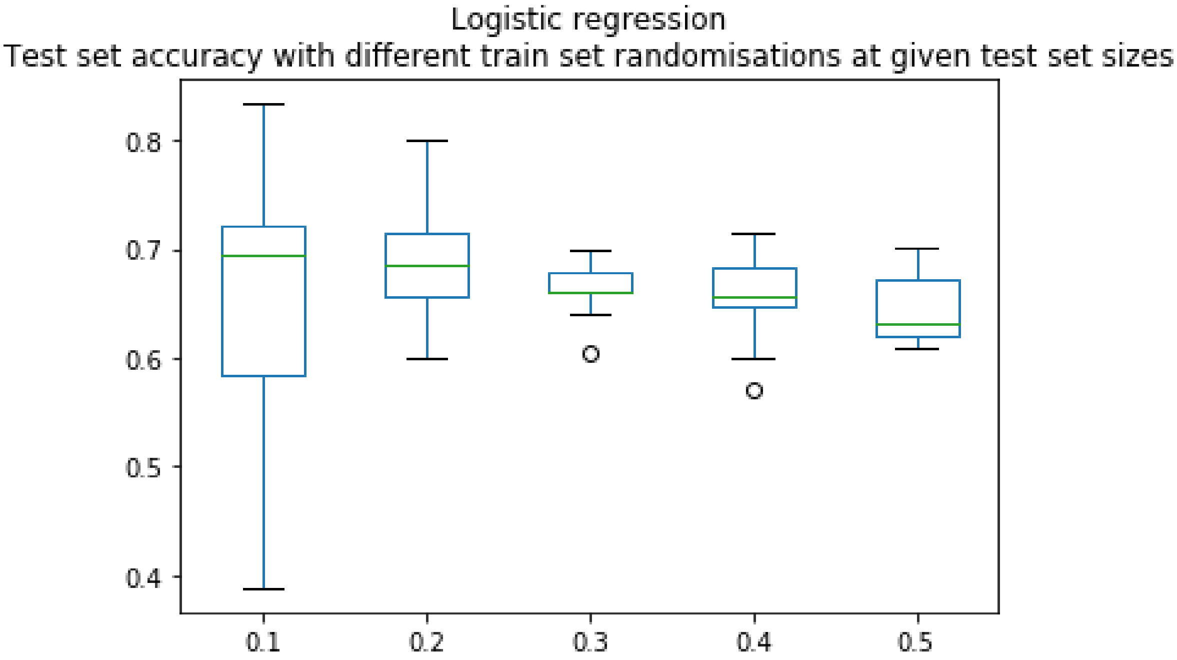
The algorithm is trained on 80% sample of the entire data set, probing different K values. As the diagram below demonstrates K=9 delivered a successful attempt to predict the labels with more than 75% accuracy, thus it was possible to find a way to identify neighbourhoods on the sides of the latte line based on their venues.



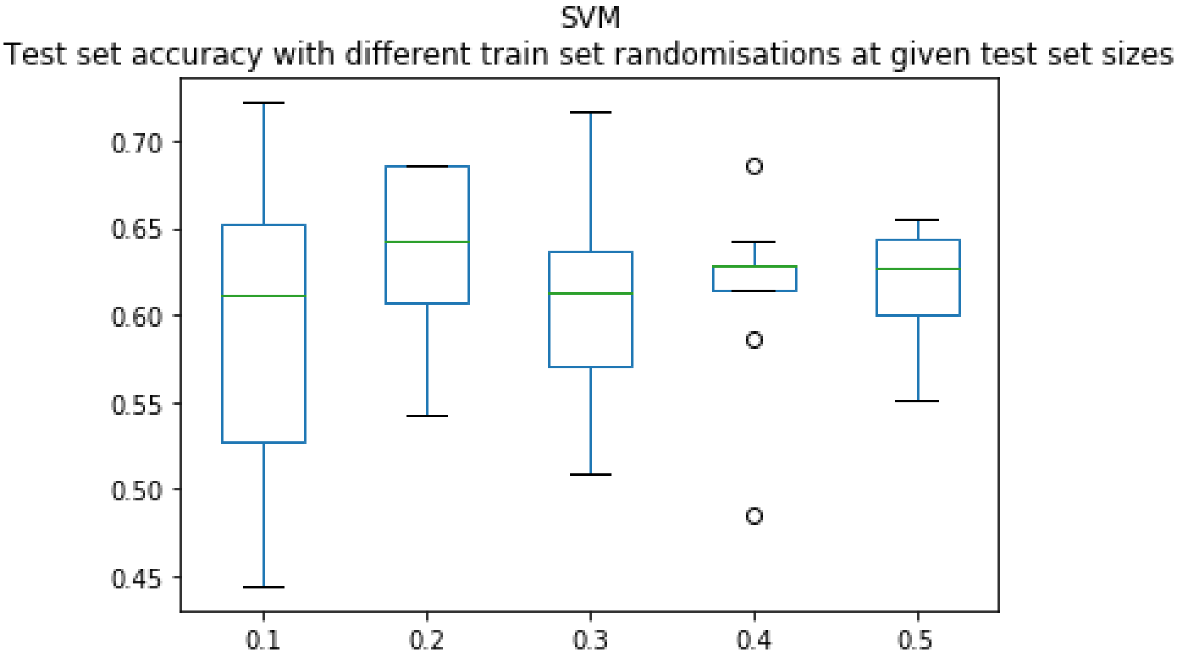
2.4.2. Logistic Regression, SVM and Decision Tree

Since our data has many different venue-types the random selection process of creating train/test sets can have a large impact on the outcome of analysis. For this reason, SVM, Decision Tree and Logistic Regression is executed 50 times with 10 different randomisation and 5 different test sizes. This solution is somewhat similar to bootstrapping, though the consideration and the implementation not the exact same. In the output for every test set size the average accuracy is calculated from the 10 random test set generation. This helps reducing the effect of randomisation and accidental overfitting.

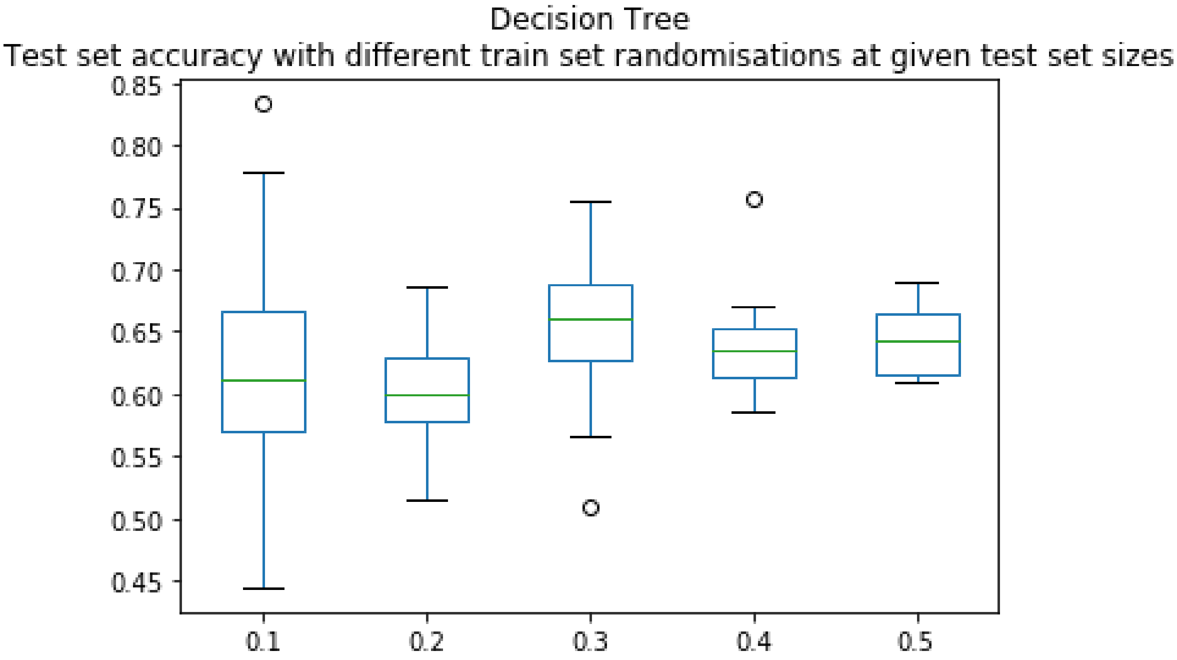
The following boxplot diagrams show the findings at different test set sizes. As seen, when the model was trained on the 70% of the data the results were consistent, otherwise the random selection had a high impact on the outcome.



SVM did not produce consistent outcome on any size of test sets.



Decision tree was relatively consistent between 20-50% but did not perform well overall.



3. Conclusion

4 machine algorithms have been tested whether with their help we can predict from Foursquare data if a suburb is north or south of the Latte Line.

Due to the low number of elements in the samples we can consider that the use of 90% of records in the training set is overfitting thus exclude their validity to test the latte line.

K Neighbours passed the initially set 75% accuracy expectations, Logistic regression was relatively close (considering the threshold was arbitrary), while SVM and Decision Tree performed below our pass mark. This suggest that based on venue types it is technically possible to predict the location of a suburb on the 2 sides of the latte line.

4. Final thoughts

It is important to note that based on venue types not only the 'Latte Line' quality can be determined since suburbs differ from and resemble to each other in many factors. It is very likely that using a different labelling e.g. to hilly and beach-side areas would also be successful, just as commercial versus residential, dense and central versus far-away and secluded. Thus, while technically it was possible to find similarities between suburbs on the same side of the line based on venue types, it does not necessarily mean that the underlying cause is lifestyle.

As analysing the best prediction results provided by the K neighbours classifier, it is difficult to agree with the classification of many suburbs if we consider lifestyle only.

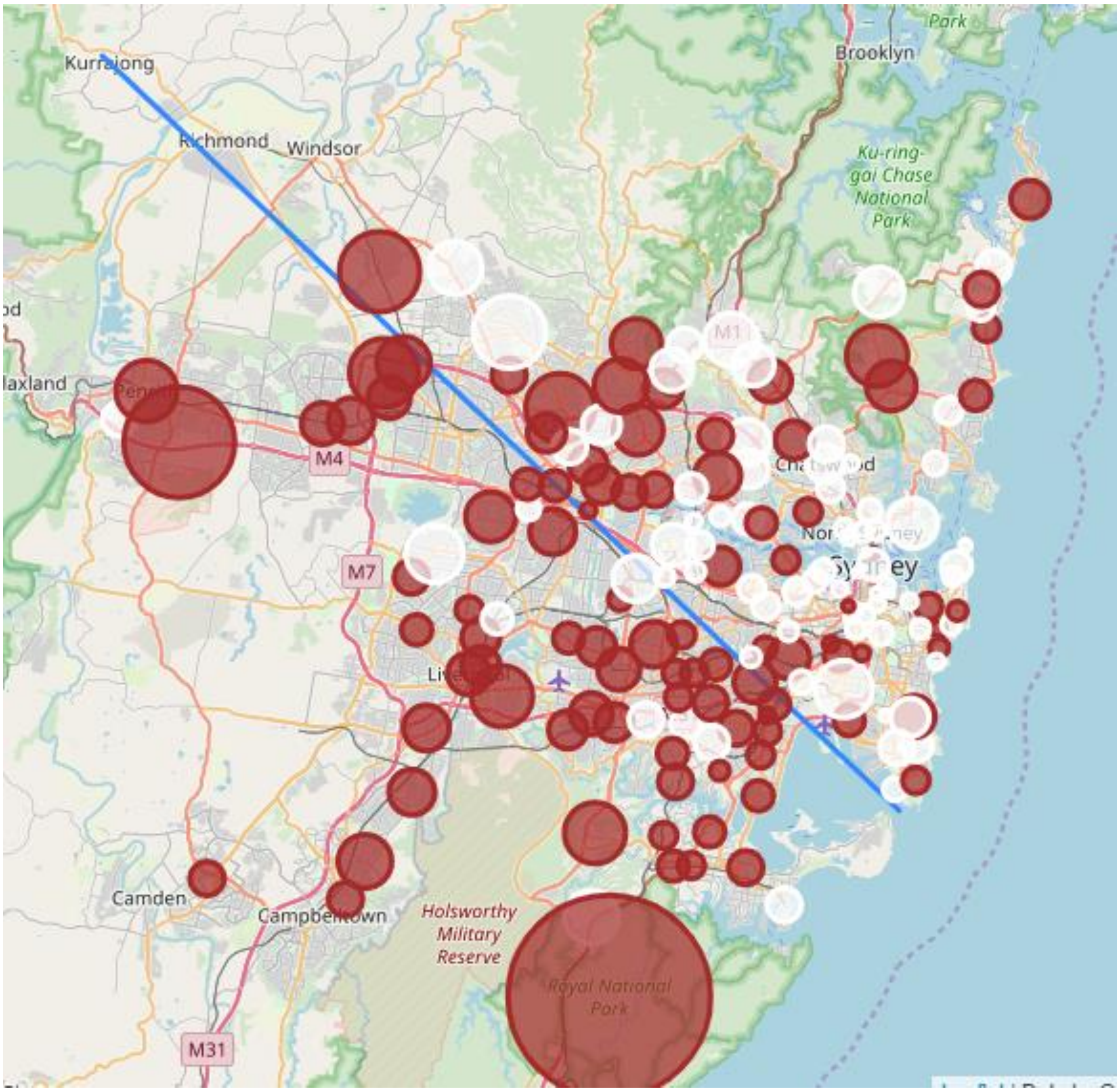


Figure 3: Predicted location of the suburbs with the best performing model