Evaluating the Reliability of the Sydney Bus Network Using Machine Learning

1 Abstract

The Sydney transport network is widely regarded as inefficient and poorly designed. With many infrastructure projects under construction such as the north west metro train line and the CBD light rail there is currently large amounts of congestion. The continuing privatisation of segments of public transport infrastructure also looms over the bus network. The opendata NSW program allows us to access a life feed of the bus network and analyse its dyanamics. We can use machine learning to investigate which factors cause lateness in the bus network and make suggestions for how to fix them.

2 Background and Research problem

We aim to determine the effects of three things on the bus network: Geography, route of the service and time of day. We expect our results to show that these factors have roughly equal importance on whether a service runs late or early. The null hypothesis for this study is simply that we will see no significant difference between routes or times of day and that geography will have a similarly negligible effect. The alternative is that these factors will have a significant effect on the lateness and reliability of these services. Traffic modelling generally expects geography and time of day to be the main factors. We will see if route factors into the analysis at all. For previous work on this topic see [1]

3 Approach Description

In the first stage of the project we collected 1.9GB of data. A line in our csv file corresponds to a stop event for a service that ran in the state of New South Wales. We collected 23, 254, 071 stop events from the 22nd of August to 21 September 2018. We collected much more data but unfortunately the majority of it was lost on a hard drive that was not backed up. Data collection has been restarted and the results will be redone in a few months time.

Early attempts to perform fits and classifications on this raw data were unsuccessful. All predictions were in the on time category. We presume that this is due to systematic under reporting of lateness from the bus drivers' consoles either due to negligence, human error or equipment malfunction. In a chaotic environment such as traffic it is unreasonable to expect many, if any stop events with 0 deviation from their timetable. So we assume that any 0 delay responses are incorrect and remove them from our set.

This leaves us with over 22 million data points. The issue did not entirely go away so we narrowed down our data set further by considering what we are truly interested in. Habitually late busses at peak and rush hour. In order to identify clusters of late busses we did a simple Kmeans clustering on time of day vs delay. This helped us identify which data points to classify and then predict. We don't want to choose routes and data that are mostly on time because in these cases lateness is too much an exception rather than the rule and so becomes difficult to predict. This is partly due to the broad definition of 'on time' that TfNSW uses. We want to discover what makes a bus late and so we need lots of late busses.

By having such a rigid, broad definition of what is considered late our algorithms have a tendency to under fit which is hard to overcome. We do not consider any routes more than 50% on time in order to avoid this problem. To supplement the random forest classifier we also do work with a gradient boosted decision trees classifier because it is less prone to underfit or over fit than the random forest. Note cutting down the data in this way only leaves us with 200 000 data points. These points are from 750 routes, more than enough for our analysis.

To choose parameters for our decision trees we take an approach common with classification algorithms. We find the best fit we can, allowing over fitting by specifying no max depth on a simple decision tree algorithm. Then we systematically cut back the decision tree and note the difference in accuracy as we do so. We then balance the computational expense and accuracy of choosing a number of estimators in our random forest. Our random forest only achieved a marginal improvement so we moved to the gradient boosted decision tree classifier with much better results.

4 Evaluation Setup

The clustering in Figure 3 is not our main analysis tool. It was simply done in order to identify which data points were most useful to us. Our main goal is to discover which deciding factors will cause a bus to be late early or on time. For this reason we will train a classification model and so use an f1 score as our metric. We will also use the feature importance coefficients to determine which factors have the greatest impact on bus lateness.

To gain a better idea of the trends within the data we looked at the geographic information in Figure 4 in the Appendix and the time information in 3. This helped us better choose our data and our classification algorithms.

5 Results and Analysis

Our initial fit was skewed towards the on time category due to poor data quality. This can be seen in Figure 6. After filtering, our fit visually improved dramatically as shown in Figures 1 and 2. However, the f1 in the random forest classifier did not show a significant increase. Only obtaining an f1-score of 0.01 higher. It could not identify the more intricate rules at play in this chaotic environment. Either hundreds or thousands more trees would have to generated or we have to increase max depth. Either would have increased over fitting.

We decided to use a different classifier, one more robust and better at picking out fine features. Hence our use of the Gradient Boosted Decision Trees classifier which can be seen in Figure 1.

Importance

Route	time of day	stop location
0.140824	0.445842	0.413334

Results

	precision	recall	f1-score	support
0	0.84	0.84	0.84	30476
1	0.86	0.87	0.86	51166
2	0.79	0.78	0.79	50184
avg / total	0.83	0.83	0.83	131826

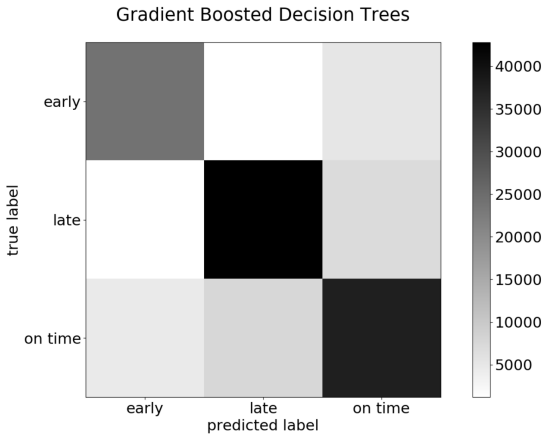

Figure 1: The results of Gradient Boosted Decision Trees Classification

Figure 1 is our main result. It demonstrates all that we had hoped to find. We can see that counter to our null hypothesis there is significant correlation between the time of day and geographical location of the bus and its lateness. While there is no significant correlation between the route of a service and the lateness. By only analysing routes that are chronically late we have not cherry picked our data, but rather we have managed to isolate the variables which play a role in bus lateness. These same factors would be found to be important on a larger data set but by training this algorithm on a small subset we can better learn which factors influence bus lateness. Our results make sense as if we look back to Figures 4 and 3 we can see that lateness is clearly a function of time of day and geography. The large red sections of Figure 4 indicate how important geography is with clusters in the city and in the northern suburbs where train services are lacking. And Figure 3 shows us that lateness fluctuates with time of day.

This forces us to conclude that the lack of reliability for busses is not actually Transport New South Wales' fault. They are unable to control traffic or geography. They can only react to it. These results simply show that the existence of lateness is predictable not that individual services can be changed to better suit traffic conditions. That would require much more data than was available for this project and it may prove impossible to improve existing timetables. These results cause us to expect bus lateness to simply model traffic. Given the resolution and scope of the data available one could in principle live track traffic across the state with the opendata protocol.

The dependence on time of day in predicting tardiness is unsurprising. We would expect when there are more services, more of them would be late. Moreover, when there are more cars on the road traffic becomes chaotic and so we get delays. However, the dependence on geography reveals something deeper. It is not a routing problem that would allow us to fix the lateness issue. It is a demand problem. So long as people wish to go to congested areas there will be an issue with scheduling in a chaotic environment. On a conceptual level this project is studying Sydney's poor city design which has been an issue since foundation.[2] The construction of more infrastructure in congested areas and the expansion of the city should hopefully curb the congestion issues and hence the lateness problem. Since we cannot control *when* people wish to travel we must control *where* they wish to travel since these are the deciding factors in tardiness. This analysis is in line with the plans of the city to expand into Paramatta and further West.[3] This will not only deal with the new wave of imigration brought about by the prosperity of the last few decades but it will also ease the demand on housing in the inner city which has caused an affordability

crisis for decades. These issues and many more are entwined in the data we have collected and analysed here.

References

[1] N. Gladstone. 'late every time': Sydney's worst bus revealed. [Online]. Available: https://www.smh.com.au/national/nsw/late-every-time-sydney-s-worst-bus-revealed-20180704-p4zphp.html

[2] T. S. M. Herald. Sydney 2026: Our green spaces. [Online]. Available: https://www.smh.com.au/interactive/2016/sydney2026/chapter5.html

[3] Three cities | greater sydney commission. [Online]. Available: https://www.greater.sydney/three-cities

Appendices

A Additional Plots

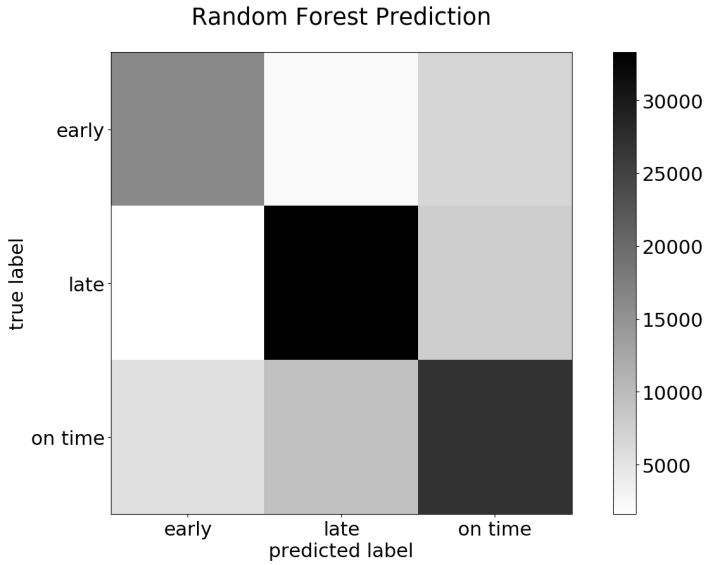

Figure 2: The confusion matrix associated with the cut down data set from the random forest classification algorithm. n_estimators = 500, max_depth = 10

```
Importance
Route        time of day    stop location
0.24339647   0.31908026     0.43752326
                  precision    recall  f1-score   support

           0       0.70      0.64      0.67     25477
           1       0.74      0.78      0.76     42609
           2       0.65      0.64      0.65     41769

   micro avg       0.70      0.70      0.70    109855
   macro avg       0.70      0.69      0.69    109855
weighted avg       0.70      0.70      0.70    109855
```

The classification report for the random forest after filtering.

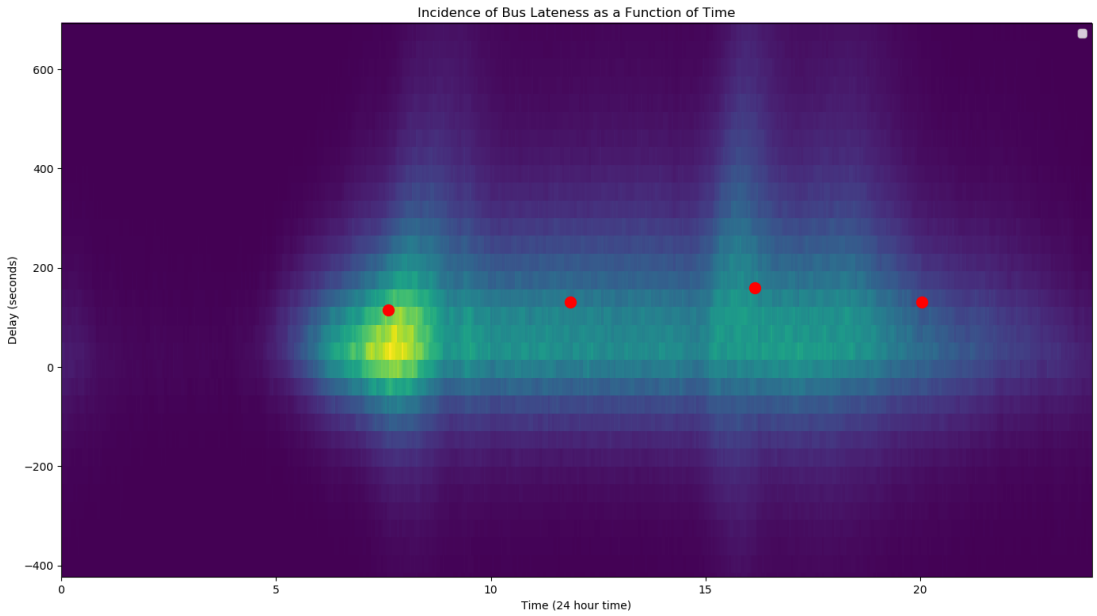

Figure 3: Aggregated data with cleaned out data points that report 0 lateness. Notice the peak periods and the tails indicating services getting later and earlier throughout peak periods. This two dimensional histogram displays 22 million data points and so it is not possible to generate the silhouettes associated with this kmeans clustering algorithm.

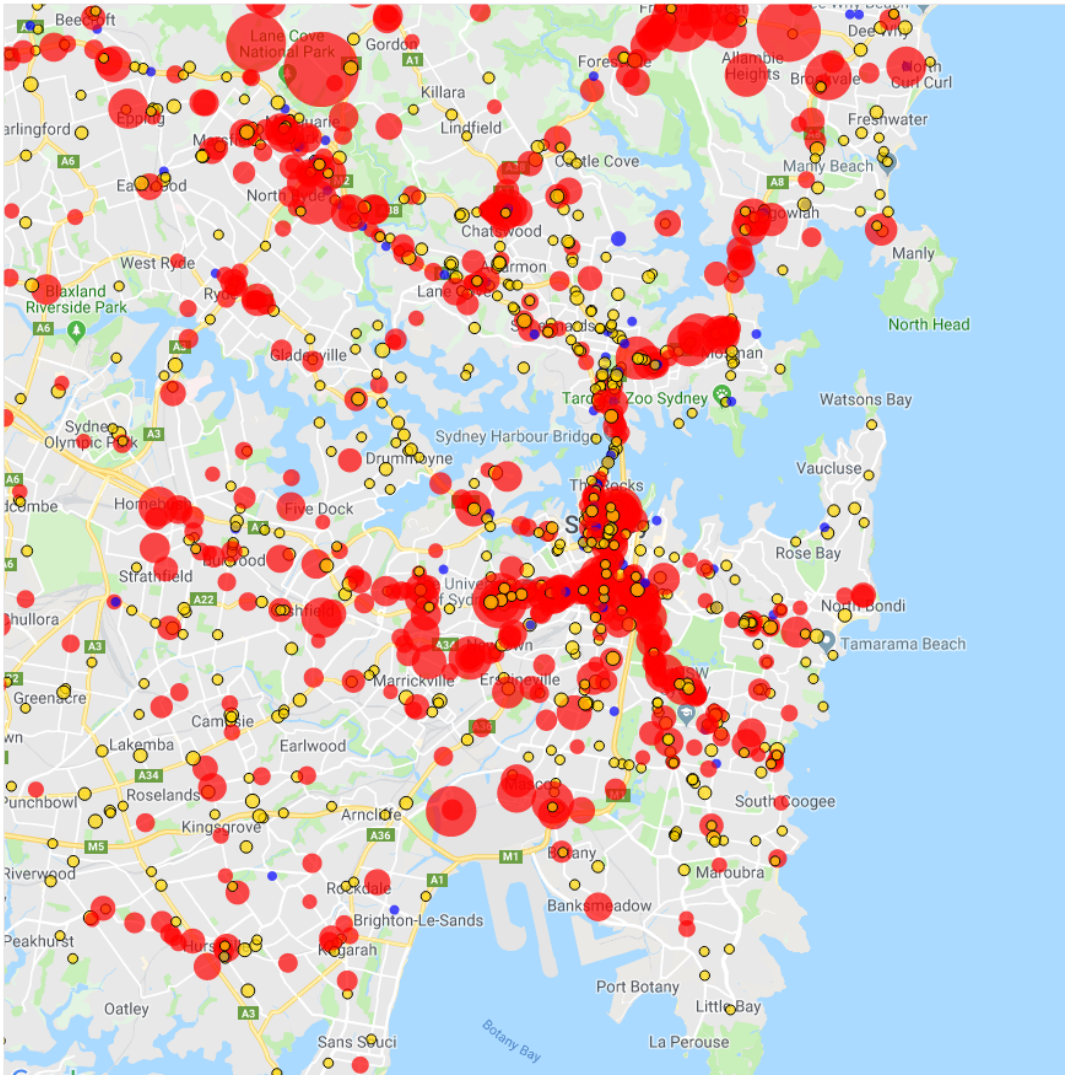

Figure 4: The distribution of lateness around the CBD at 9:45 am October 23rd. This helped us better understand the importance of geography in the network. Gold indicates a service is on time, blue indicates early and red indicates late. The size of the point indicates how late or early a service is.

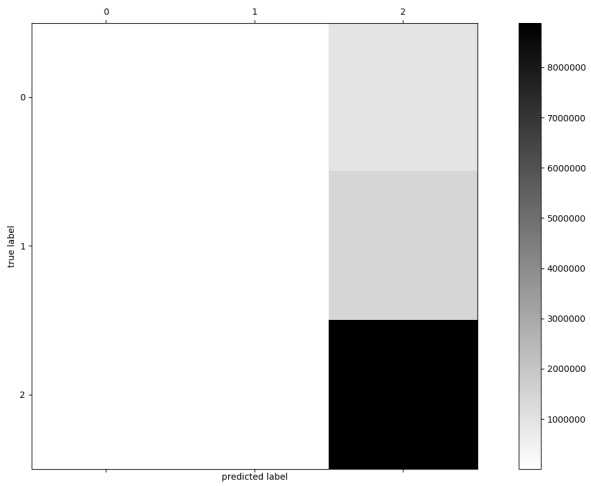

Figure 5: The poor fit by the initial random forest matrix

	precision	recall	f1-score	support
0	0.69	0.00	0.01	958206
1	0.57	0.00	0.00	1439284
2	0.79	1.00	0.88	8890565
avg / total	0.75	0.79	0.69	11288055

Figure 6: The classification report of the random forest. Note the poor result. Not much better than random.

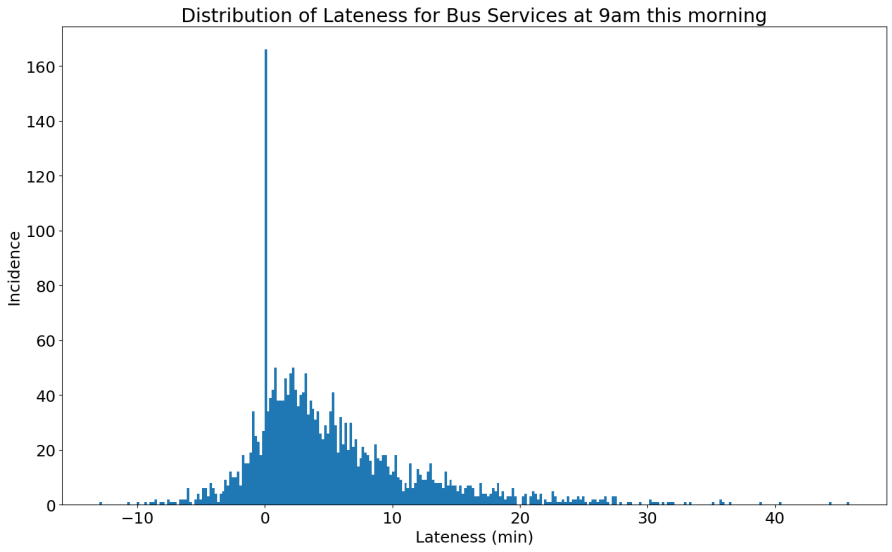

Figure 7: The incidence of delays accross the network at 9:45 am on October 23rd. Note the anomolous 0 readings and close adherence to a gaussian fit.