Citadel Datathon Dublin - 2017
Section 1: Topic Question

In what ways can areas sensitive to targeted advertising by Uber be identified, and how have the predictors for these areas evolved over time?

How to target users is one of the great challenges of advertising, and uber trip data combined with standard taxi data provided a perfect opportunity to explore this. (See: Google now)

The goal of our project is to identify the conditions - namely, location, day of the week, time of day, month, weather conditions - under which people are very likely to either use exclusively a taxi or Uber. We make the assumption that such users are unlikely to switch from using a taxi to using Uber and vice versa. At the same time, we identify the conditions under which people use both Uber and taxis. Under these conditions, it is possible that clever advertising in the right place and at the right time could be effective (for example, via notifications on people's mobile phones when such conditions are met). Time permitting, we also aim to explore how these conditions evolve over time, as well as verify our hypothesis that under certain conditions people are likely to continue using either Uber or taxis almost exclusively.

In our Analysis we considered the following datasets:

- Uber trips (2014 & 2015)
- Yellow cab trips (2014 & 2015)
- Green cab trips (2014 & 2015)
- Geographic
- Weather
- Zones

Section 2: Non-technical Executive Summary

It is infeasible to process the amount of raw data available on a typical laptop in the timeframe allowed for the competition. Therefore, early on a decision was made to randomly sample 1% of the trip data and to use the sampled data in our data analysis. The same analysis could then be run on the full data, given enough time.

The first task was to combine the data into a single dataset that could be easily analyzed. Each row in our dataset was to correspond to a single taxi or Uber trip, and the columns were to be as follows:

isuber : Binary flag indicating whether the trip was using Uber or a conventional (green or yellow) taxi
pickup_datetime : when the passenger was picked up
pickup_latitude : where the passenger was picked up
pickup_longitude   where the passenger was picked up
max_temp : maximum temperature on the day
min_temp : minimum temperature on the day
avg_temp : average temperature on the day
precipitation : precipitation on the day
snowfall : snowfall on the day
snow_depth : snow depth on the day
month : month index, since Jan 2014 (integer between 1 and 24)
hour : hour of the day, between 0 and 24
DayOfWeek : day of week, between 0 and 6

Uber trips data for 2015 did not contain longitude and latitude for the pickup locations. We had to recover that data from the zones and geographic table. For simplicity, we took the mean of the vertices of the region boundaries to impute the pickup locations. This is a very crude imputation technique; with enough time, this could be refined to sampling pickup locations from the regions according to the real pickup location distributions within those regions, which could be recovered from the Uber 2014 data.

Having performed a quick initial analysis of the data, we identified that some features may have some predictive power for whether a user was likely to take a taxi or an uber (namely: location, temperature, time of day, day of week - pair plot is included in Appendix)

This formed the basis for our solution. As opposed to trying to simply predict whether a user was likely to take a taxi, we aimed to explore the areas in which a classifier loses its predictive power; where the distribution of transport choice was unaffected by surrounding conditions.

In turn, this may imply regions where users have no clear motive to choose between the two transportation methods, and be best swayed by targeted advertising.

Our model, described in the technical summary, was trained to predict whether a user would choose Uber based on a set of engineered features. It achieved an initial Cross-validation accuracy of 0.93; it is important to bear in mind that this model was trained purely to explore the area in which trips were difficult to predict.

Once we had a model that could differentiate between certain Uber and taxi journeys, we explored the boundary at which the model did not produce a clear classification. Statistical tests based on samples in this area showed very positive results. The statistical similarity of feature distributions in this boundary region was high - indicating there were no apparent differences between uber and taxi travellers - therefore a potentially sensitive area to marketing. These results are below, and demonstrated that where all features were previously from different distributions, in the boundary zone several features now appeared to be sampled from the same distribution - indicating a potential increase in marketing sensitivity.

Statistical testing

- We statistically tested whether the features were drawn from the same distribution both for the **whole sample** and for the **boundary cases**.
- For the **whole sample** we found that **all** features were tested to be from **different** distributions.
- The results for the latter are summarised in the following table:

Feature	**p-value**	**Are the samples drawn from different distributions at the 10% significance level?**
Pickup Latitude	0.0002	Yes
Pickup Longitude	0	Yes
Max Temperature	0.03	Yes
Min Temperature	0.55	No
Average Temperature	0.14	No
Month	0.03	Yes
Year	1	No
Hour	0.28	No
Day of Week	0.48	No

- In other words for the **minimum temperature**, **average temperature**, **year**, **hour**, and **Day of week** are said to be drawn from the same distribution.

- Conversely, the **boundary cases** the **coordinates**, **maximum temperature**, and **month** are all said to be drawn from different distributions for the Uber and non-Uber cases (as with the **full sample**).

For more details on the statistical tests conducted please refer to the relevant part of the technical section.

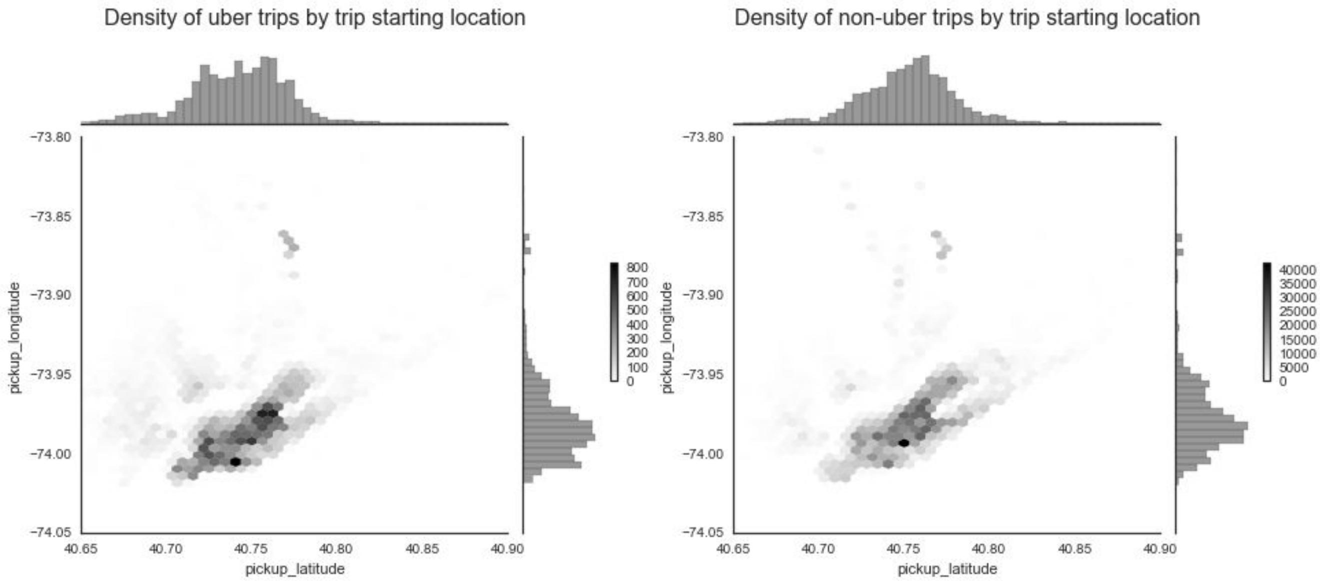
Section 3: Technical Executive Summary

The overall process followed to create the sensitivity model was as below:

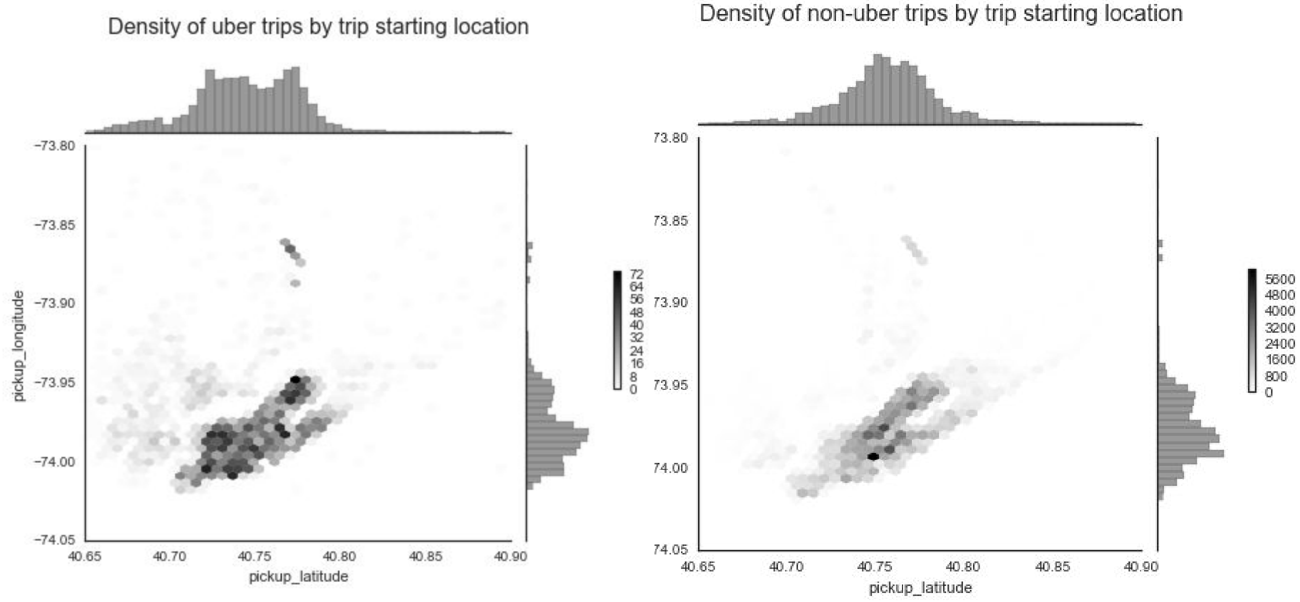
- Initial Data Analysis
- Dataset preparation
- Coordinate parsing
- Model creation

How these tasks were approached is described below, followed by a more in depth analysis of our results.

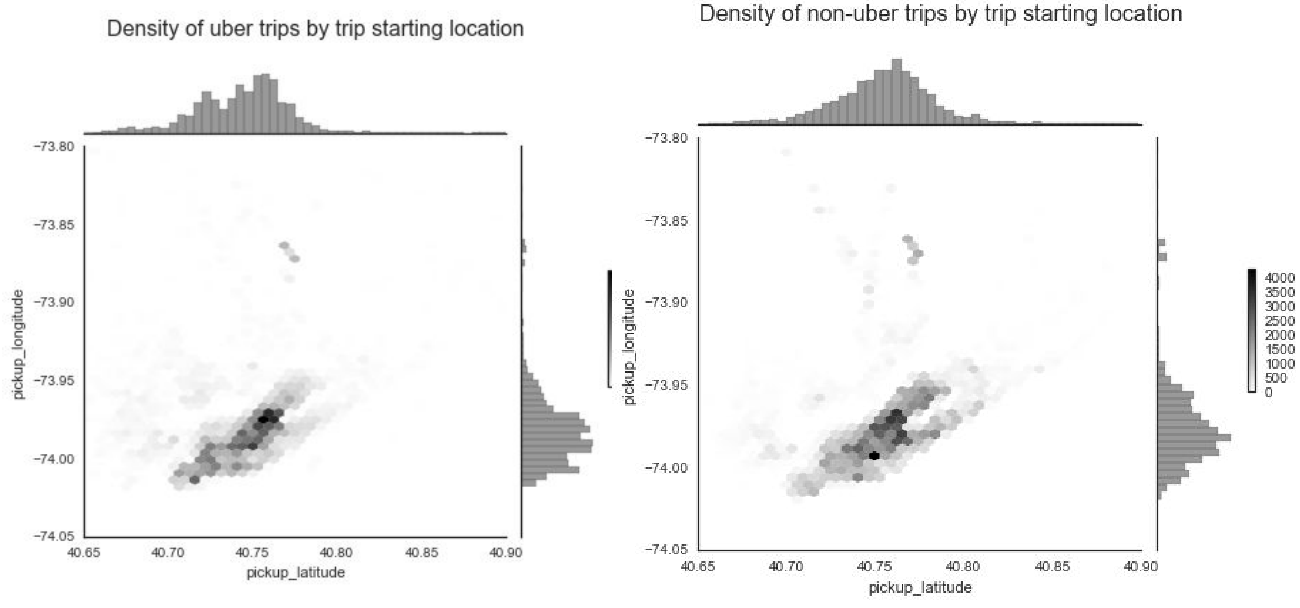
Initial Data Analysis



Above: Plots showing geographical distribution
of uber and non-uber rides starting locations



Plots showing locations between 6am and 10am,
likely commuter traffic



Plots showing pickup locations between 5pm and 7pm,
likely commuter traffic returning home

We began by examining the differences in trip data and uber data geographically. The most notable comparisons to be made were the differences in travel patterns between morning and evening journeys - with many commuting into the city in the morning, and from the city to the suburbs in the evening.

Dataset Preparation

The dataset was initially highly fragmented, and most time was spent here deciding how to best approach combining this into a single set. Eventually the process followed was decided as below.

- Downsampling all data to 1% of its original size, as described above.
- Producing labels for every journey - either Uber, or Taxi. It is of note that green and yellow taxis were included under the same label.

Coordinate Parsing

In 2015, Uber changed its location format to be parameterised by taxi zone ID rather than coordinates. This means we had to map from this to coordinates to be consistent with the rest of the data. We mapped the taxi zone ID to NTA code in the **Zones** dataset, and then mapped from this to corresponding coordinates of the polygon in the **Geographic** dataset. To arrive at a final set of coordinates for the journey, we simply averaged over the coordinates relevant to the polygon.

Data Leakage

It is worth noting here that all 2015 uber trips would only have a set of 263 latitude or longitudes. This introduces potential data leakage, which a highly nonlinear model could exploit to boost its prediction accuracy. We did not have time to explore this any further, however given longer, sampling noise to add to these results is something we would have liked to explore.

Machine Learning

While it would have been interesting to explore more complicated options, we used a random forest estimator to predict whether a journey was likely to be uber or taxi. This was trained on the sampled dataset, and validated on a dataset with an ~0.33 split size.

Note also that the weather was given as a function of location and time of day. For simplicity we focused on only the time when considering the weather, i.e. we used the same location throughout. This corresponded to the weather at JFK Airport.

Statistical testing details

Below we discuss some of the technical details of how the statistical testing was conducted:

- To perform statistical analysis on the feature values between the Uber and non-Uber distributions in the boundary zone, we perform a **two sample Kolmogorov-Smirnov test**, where the null hypothesis says that two empirical samples come from the same distribution.
- If the test is rejected it means that the samples have been drawn from different distributions (although it says nothing about the behaviour of these).

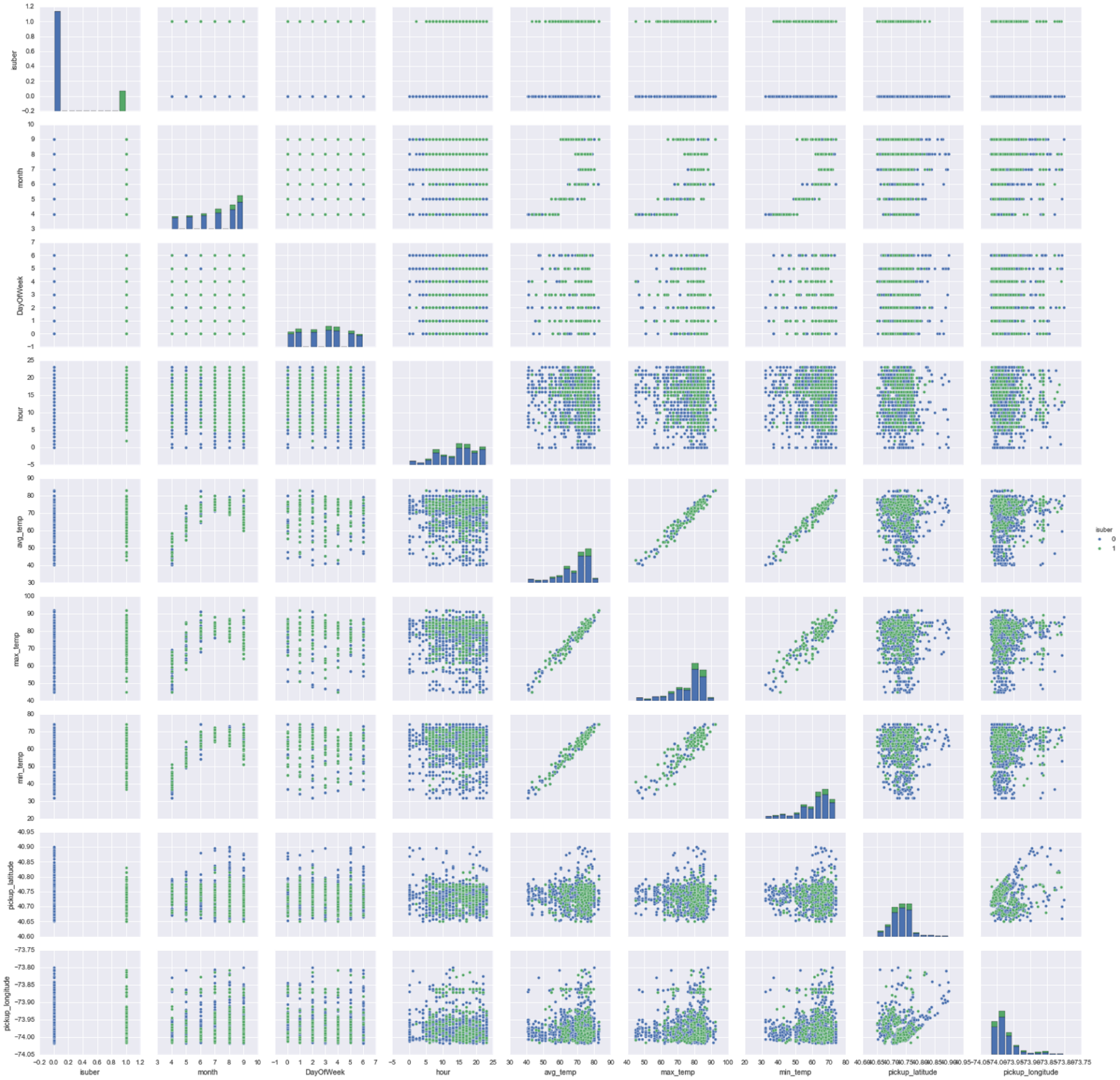
- We calculated the **two sided p-value** from the test, and compared it at the **10% significance level**.

Final Conclusions

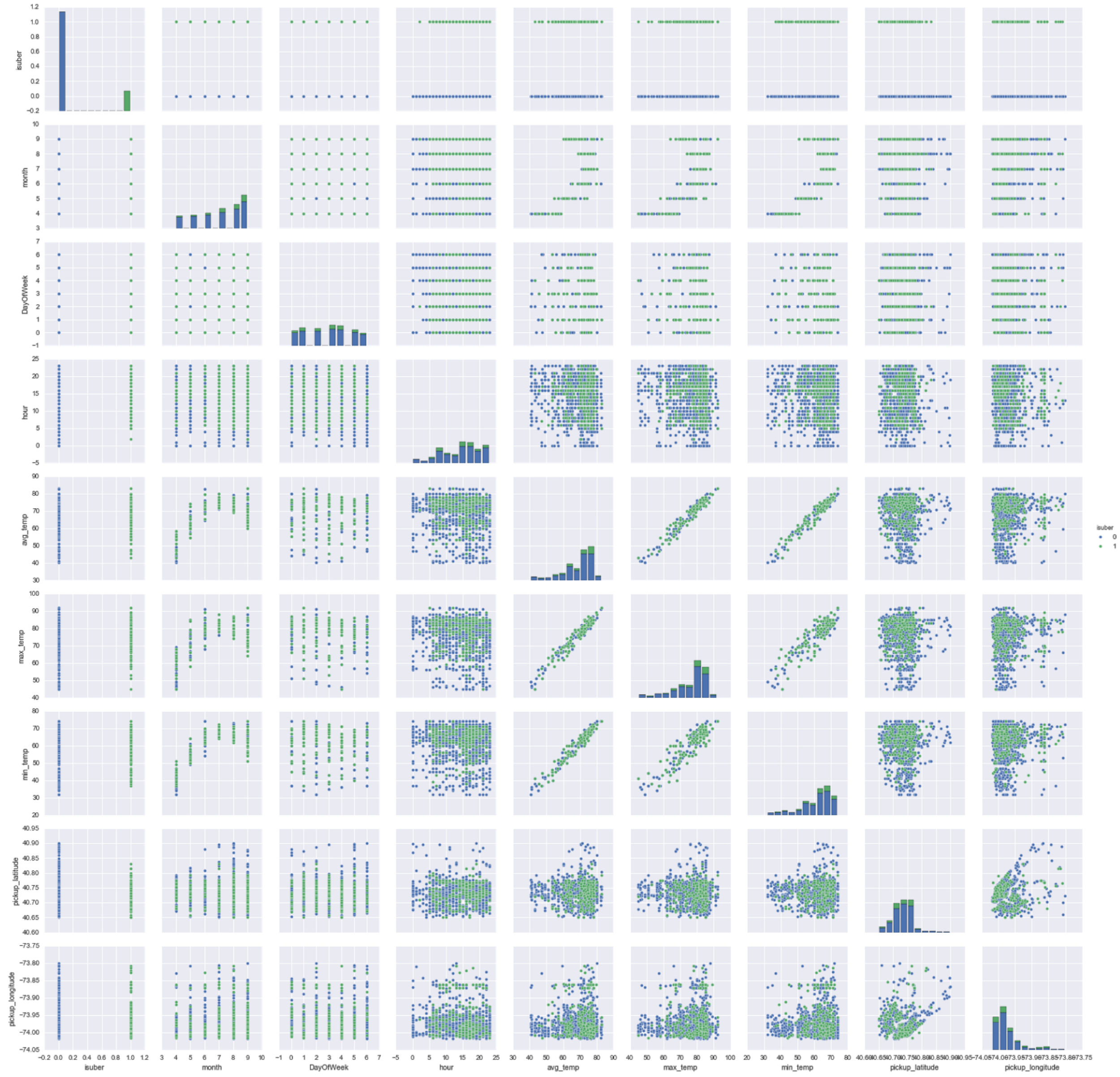
The most important output of the model we trained is whether the samples on its decision boundary were statistically similar. We compared this in two ways - firstly the per-feature statistical tests performed in the non technical summary, and secondly by visual inspection of new pair plots, comparing them to the pair plot of the original data. The statistical tests, as described earlier, indicated higher similarities between the distributions, and the pair plots show much more closely clustered data. While much of the reasoning behind this targeting method is empirical, it provides an interesting metric as to how marketing sensitivity could be approximated, to a degree higher than human decision trees.

Appendix

Initial Analysis Pair Plot



Boundary Zone Pair plot


isuber
month
DayOfWeek
hour
avg_temp
max_temp
min_temp
pickup_latitude
pickup_longitude
isuber