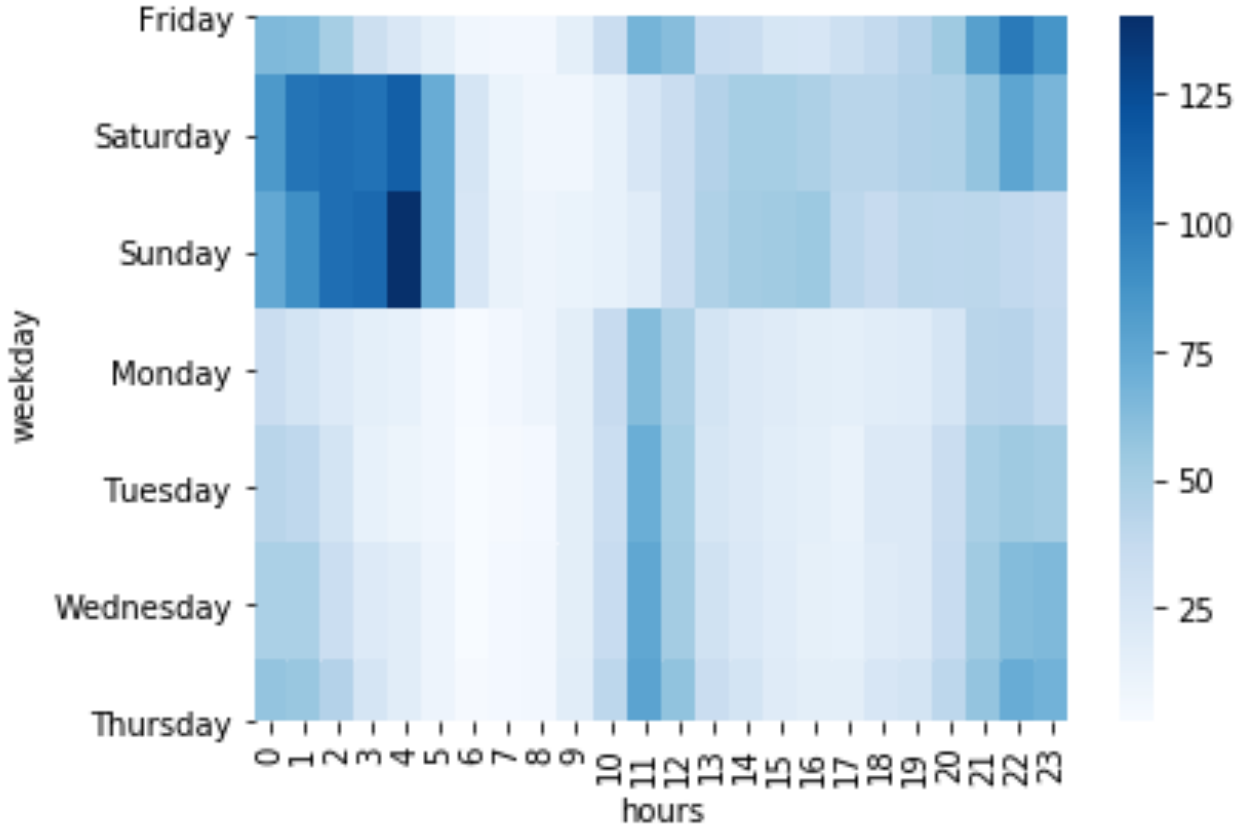
Ultimate Takehome Interview Challenge

I. Exploratory Data Analysis

The heatmap set forth below (and the table underneath it provided for numerical reference) illustrates the underlying patterns of demand on an hourly basis for each day of the week. From this it is strikingly clear that the peak hours occur in the early hours of Saturday and Sunday morning, specifically between 1am and 4 pm, when the average number of trips generally exceeds 100 per hour. Friday night at 10pm is also a peak. One may speculate that users are taking cars to avoid possible drunk-driving problems.

There are also much more modest peaks during the working week before noon, possibly related to eating out at restaurants, or running personal errands.



hours	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23
weekday																					
Friday	65	64	51	33	24	16	8	7	7	16	...	35	26	25	32	38	44	54	80	101	87
Saturday	84	104	107	105	115	73	27	12	8	8	...	51	51	48	43	43	46	47	58	77	67
Sunday	75	90	107	110	140	73	25	13	10	12	...	52	53	55	41	37	42	41	42	39	37
Monday	35	28	21	16	14	8	4	7	10	17	...	22	20	17	16	18	19	27	43	44	38
Tuesday	43	40	28	14	11	8	3	5	6	17	...	22	18	16	13	22	22	35	50	54	52
Wednesday	49	49	34	21	18	10	3	6	7	17	...	23	19	15	14	20	22	36	53	63	65
Thursday	58	56	45	27	18	10	5	6	7	18	...	28	20	17	17	25	28	41	58	73	69

II. Experiment and Metrics Design

1. The business metric that I would ideally select would be whether the increase in revenues to Ultimate exceeded the aggregate costs of the tolls paid by Ultimate.
2. A crude proxy for this might be the aggregate increase in weekly-miles driven by all drivers.
3. A more precise benchmark would be to focus specifically on the drivers for which a toll was paid, and to sum the miles for their trips (on those days that they had paid the toll).
4. We can then make calculations to support a statement of the following form: *"Of the drivers who on any given day took advantage of the Toll Offer, their aggregate miles driven was x: for the same group of drivers, normally their aggregate miles driven was y".*
5. The most statistically "powerful" approach would be a so-called one-tailed "*Paired t-Test*", in which for each driver a direct comparison is made. The standard deviation for calculating the z-score is that of the *differences* for each of the various drivers. Bootstrapping is another way in which statistical significance could be calculated via a p-value (with no assumptions about underlying distributions). An important advantage of having a "powerful" test is that statistically significant results can be obtained from a much smaller sample size.

Sizing the Offer: A somewhat subtle business point here is that there may be so-called diminishing marginal returns to Ultimate from their offer. In other words, it might be attractive to offer it to a fraction of drivers, but NOT to all of them. Repeated A/B tests at different times of the year and of different sizes (eg respectively 10%, 20%, 30%, and 40% of drivers) would help quantify this. The goal would be to identify the threshold level at which the market became "flooded", so that beyond this point the incremental miles driven were insufficient to have warranted the offer of toll reimbursement. This threshold level is likely to vary significantly with the hour of the day, and day of the week, and so the "*Heatmap Approach*" outlined immediately below would appear to be particularly constructive. Seasonality is important, as the heatmap for the Holiday season in December may be much different than for January, for example.

Heatmap Approach

An even more refined approach would be *for those days that the toll was paid*, calculate a heatmap/table of hours driven, and then subtract from this the baseline heatmap/table for each respective driver, to obtain the differences. By summing the tables of differences, an aggregate table of differences can be obtained, from which a summary heatmap can be generated. The "Paired t-Test" approach can then be used element-by-element for the differences map for each driver. So effectively the test would be repeated 24 x 7 = 168 times.

This heatmap might be helpful in identifying whether there are particular hours when the offer might be made, as opposed to all hours of the day. For example, based on the heat map in ***Part I*** above, it might be optimal to make the toll-reimbursement offer just for the early hours from midnight to 4am for Saturday and Sunday mornings.

In terms of implementation as an A/B test, a subset of drivers should be selected at random. One question is whether on a long term the subsidy can be offered to a subset of all drivers, or whether this might be regarded as potentially discriminatory.

Drivers obviously the driver needs to record whether the toll was paid to be eligible for reimbursement, and ideally the time at which it was paid.

III. Predictive Modelling

Using the Scikit-learn machine learning library for the Python programming language, a variety of Classification models were constructed to seek to:

(a) predict rider *retention* for Ultimate's users six months from signing up,
(b) to help understand what factors are the best predictors of retention, and
(c) to explore how the insights gained might be operationalized to help Ultimate.

A user is considered retained if they were active (ie took a trip) in the preceding 30 days. Since the latest "*last_trip_date*" timestamp available for the cohort is June 30th, 2014, any rider who had taken a ride in the thirty days of June was considered to be "active". By this metric, 36.6% of the users who had signed up six months prior were found to be 'active'.

A dataset was available for 50,000 users who had signed up from three cities only, with ten variables in addition to the sign-up date, and the last date when a ride had been taken.

Regarding the NaN's for driver rating, I have speculated that the reason for the non-rating may have been an element of dissatisfaction: accordingly, the NaN has been replaced with a value corresponding to the lowest quartile. Normalization was not undertaken, as Decision Tree models are not effected by normalization [but KNN?].

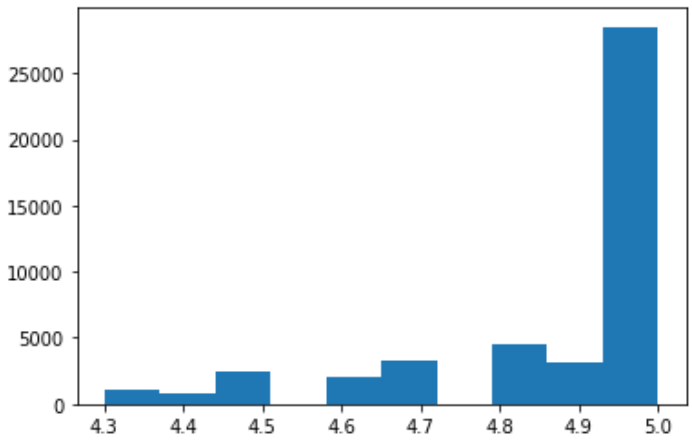
What factors are the best predictors of retention?

A variety of models were developed and their performance was evaluated. The overall performance of the models was quite similar, with accuracy of around 73%, and an AUC (*Area Under the [ROC] Curve of 0.72*). None of these models would be regarded as highly predictive, and their relatively weak performance might be because many important variables (such as for each user their age, gender, ethnicity, location, and income associated with their zip code) were not considered. The models included:

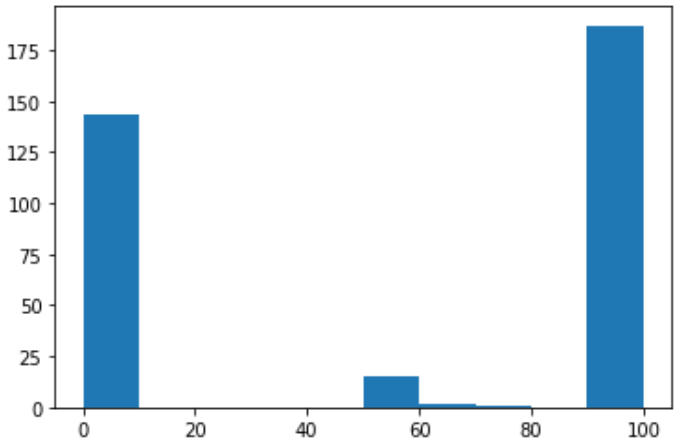
1. Decision Tree
2. Random Forest
3. Logistic Regression
4. K-nearest neighbors
5. Naïve Bayes
6. Support Vector Machine

However, working with the data available, the model primarily focused on was Logistic Regression, because unlike the other models the results can be readily interpreted, and the statistical significance of each variable can be assessed. Specifically, the regression coefficients for the variables can be exponentiated to derive factors that can indicate how the probability of retention might change. Initially the scikit-learn logistic regression model was used, but this was then succeeded by the statsmodel version, which provides measures of statistical significance for the various parameters. Accordingly, two statistically insignificant variables were dropped (*driver_rating* and *surge*), and the coefficients were recalculated for the remaining variables. As it turned out, perhaps because of an usual lack of multicollinearity (the independent variables being mostly uncorrelated) the magnitude of the coefficients did not change much.
Regarding the individual variables, considerable time was spent examining the variable 'passenger rating', which is very heavily skewed to the left (below – LHS), to an extreme degree that pre-processing cannot obviously remediate. The figures below exclude "undesirable" users with a rating below 4.3.

Histogram of the Feature 'passenger_rating'



Bimodal Distribution of 'Weekend Use' for Users with Very Low 'passenger_rating' (below 3.0)



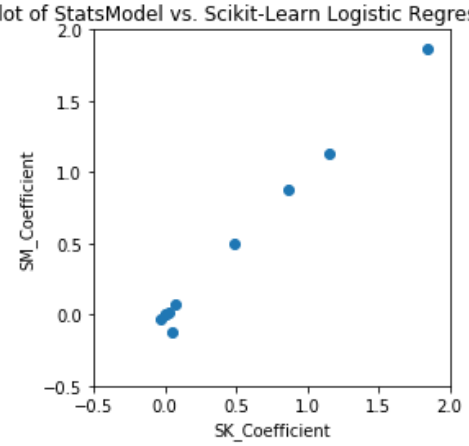
For passenger/users for whom this rating was below 4.3 (out of the maximum of 5), the retention-rate was only 24% (and only 20% for even lower ratings). One may speculate that users with a passenger rating (by the driver) of below this threshold may have had longer waits for drivers, and decided to no longer use the service. A large fraction of these users were exclusively weekend (see above – RHS figure), and may have been using the service in the "wee hours", as was apparent very common from the heatmap data evaluated on Section I above. The business interpretation is complex. One view is that these users received low ratings for their poor behavior, in terms of being late for the car, and being drunk and possibly abusive. These users might therefore be deemed undesirable by drivers. Accordingly, the data was segmented into two groups, one for users with a passenger rating below 4.3 (corresponding to the lowest x% of users), and the other (denoted *Desirable Users*) with a rating of 4.3 or above.

The overall retention rate for the Desirable Users group was 38%, for the baseline city of Astapor. However, all other things being equal, the retention rate was associated with an increase to 50% for the city of Winterfel, and to 80% for the city of King's Landing. The reasons are unclear. Could it be related to income or some other omitted variable? Or could it be related to an expensive toll or some other consideration that might cause the waits for a car to be longer in the base city of Astapor, such that users discontinued using the service.
The two binary variables, iPhone ownership (relative to the base of Android) and being an Ultimate Black user, were respectively associated with increases in Retention from the base 38% to 65% and 59% respectively. One may speculate that each of these variables may be associated with an omitted variable of income level, that might in terminology of statistical inference be deemed "confounding".

'*Revised Prob*' is Relative to the Base Retention Level of 38%: SK and SM Coefficients/Params are Quite Similar

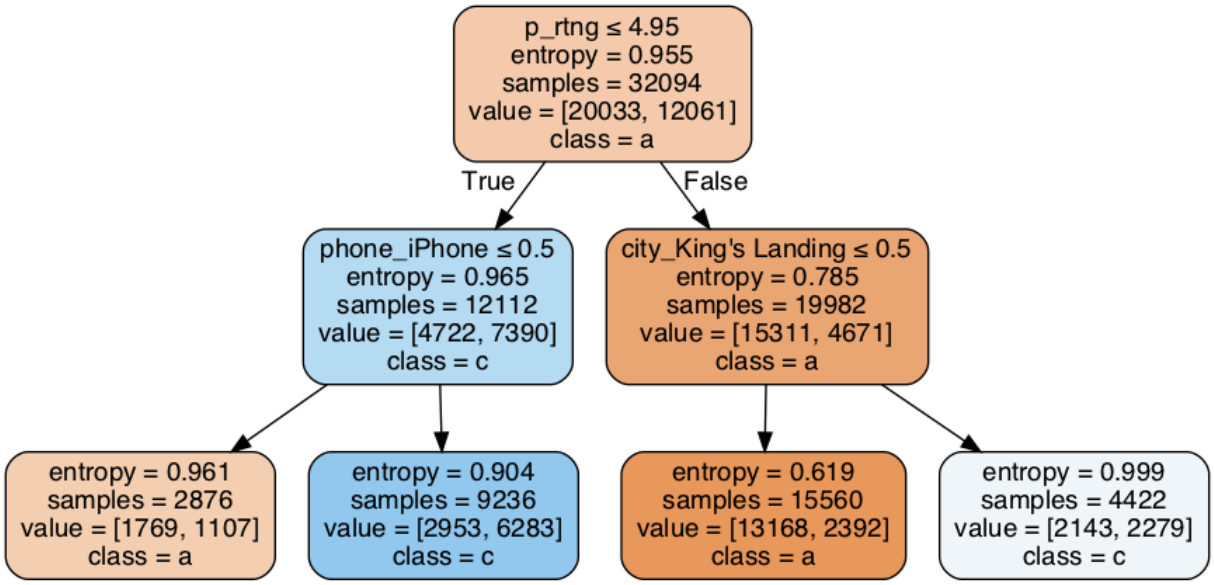
	Feature	SK_Coefficient	SK_odds	SM_params	revised_prob
0	dist	-0.031725	0.968773	-0.030494	0.371
1	p_rtng	-2.511761	0.081125	-2.759520	0.037
2	d_rtng	0.026861	1.027225	0.015929	0.382
3	surge	0.046911	1.048029	-0.122145	0.350
4	surge_pct	0.002421	1.002424	0.004334	0.379
5	f_trips	0.077419	1.080495	0.078504	0.397
6	wkday_pct	0.001237	1.001237	0.000988	0.378
7	city_King's Landing	1.840154	6.297510	1.868511	0.797
8	city_Winterfell	0.489887	1.632132	0.497090	0.500
9	phone_iPhone	1.149447	3.156446	1.135089	0.654
10	b_user_True	0.871789	2.391184	0.875369	0.593

Scatter Plot of StatsModel vs. Scikit-Learn Logistic Regression Coefficents



A Highly Speculative Thought

Although the primary model was Logistic Regression, the Decision Tree approach (which effectively focuses on combinations of factors) surfaced a rather counterintuitive observation. Namely that of the twelve variables (including categorical variables), the ***single most important factor*** in predicting whether a User would be retained was whether their passenger rating was ***below*** 4.95. This is all the more surprising given that, as already noted above, lower ratings (eg below 4.3) were generally associated with much lower retention rates (eg below 24% versus the average of c.37%). The second most important factor was whether the User owned an iPhone (which ownership might be viewed as a proxy for a higher income level). In the diagram excerpted below, the class "c" corresponds to Retention, and class "a" with non-retention.



Moreover, this is not an anomaly of Decision Trees, as ***in the Logistic Regression model the coefficient for passenger rating is strongly negative, in an amount exceeding 40 standard errors***: the odds of this occurring by chance are longer than a million-to-one.

A little online research may have pinpointed a reason why a rating above 4.95 (achieved by more than 25% of users) might be a negative. In order to obtain a passenger rating of the highest level, namely '5', it was in an article suggested that a $5 cash tip be given, even though this cash tipping practice is contrary to Ultimate's [Uber's] stated business model. This tipping practice has apparently become so pervasive that it even has its own nickname, "*Five for Five*" (so a $5 "secret" cash tip to secure a perfect passenger rating of "5").

In the bottom leaf colored in intense blue (to convey its relatively high "information value"), the retention rate for the training set was 6283/(6283+2953), or 68%, versus the overall rate of 12,061/(12,061 + 20,033), or 38%. This intense blue leaf corresponds to a passenger rating of ***below*** 4.95 ***and*** iPhone ownership. In other words, are the happiest (and most likely to be retained) users, those who are higher income (being premium iPhone owners), but who do NOT slavishly follow the "*Five for Five*" tipping practice.

Conversely, one might argue that the persistence of drivers in trying to coerce users into tipping $5 might be the single most important factor in adversely reducing passenger retention. This would be an excellent example of behavioral economics at work. Drivers are effectively acting in concert to create a two-tiered system in which they collectively identify those users who tip generously (with a perfect 5.0 rating), who will also tend to be served first by other drivers (who also seek to benefit from a potential $5 tip). This practice is strongly in

the economic interests of the drivers (who love the $5 tips), but strongly contrary to the economic interests of Ultimate/Uber, who loses disaffected riders as a result of it, and of course do not share in the tips. It might make sense for Ultimate/Uber to make efforts to enforce its stated policy, and discourage drivers from soliciting tips in this manner.

Interviews of users with ratings below 4.95 might help clarify whether this speculation has any validity. Of this speculation should happen to have any validity, it would be an amusing illustration of the complexities of human behavior patterns in the real world. Behavioral Economics has become a popular area of study, as it deals with the subtle complexities of the real world.

Post Script

I was sufficiently intrigued by this perplexing issue (vis "*A Highly Speculative Thought*" above) that I followed up with Paul Oyer, a Stanford GSB Professor and occasional Uber driver (as part of his research), whose presentation on Uber I had attended at an alumni event two years ago. His response appears to confirm this thought. It is interesting to me that I was so reluctant to accept what the models were telling me! Sometimes our "prior beliefs" can sometimes be so strong that contrary information tends to be rejected out of hand!

Incidentally, this could lead to an entirely different formulation of a Data Science question, to be addressed with the same data set:

"Uber is concerned that its drivers are rating passengers a '5' only if they receive a $5 tip (a practice so common it even has a name: "Five-for-Five". Drivers can thereby recognize who the good tippers are by their '5' passenger rating, and these users will tend to be picked up first, in effect creating a two-tiered system, that is in the interests of the drivers, but perhaps not in the interests of Uber. Using the data available, how might this question be investigated?"

-------- Original message --------
From: Paul Oyer <pauloyer@stanford.edu>
Date: 09/10/2019 8:55 PM (GMT-05:00)
To: David Willson <dwillson@stanbridgecapital.com>
Subject: Re: [Paul Oyer] Uber Passenger Retention, Passenger Rating, and Tipping

Good analysis, David! Thanks for your note.

—Paul

On Tue, Sep 10, 2019 at 6:27 AM David W Willson <dwillson@stanbridgecapital.com> wrote:

> This email was sent using a contact form at http://gsb.stanford.edu. referred from: faculty-research/faculty/paul-oyer
>
> Hi Paul, I am GSB '82 and enjoyed your great presentation at the Alumni event. These days I am learning something about data science, and in that context came across a takehome challenge by Uber, basically to predict Rider Retention. I was intrigued to see that the most important variable (apart from host city) seemed to be the passenger rating by the driver. Moreover, a rating of a perfect 5 was associated with a LOWER rate of retention than one of 4.8 or below. My wild speculation is that to get that 5 rating, one often needs to tip (even has a name, $5 for 5). Users who do not tip and have a lower rating seem more likely statistically to continue the service. I am pleased to email you my short analysis if you should happen to be interested! Your thoughts? Cheers, David Willson '82

--
Paul Oyer
Mary and Rankine Van Anda Entrepreneurial Professor
Stanford University, Graduate School of Business
(650) 736-1047 (office)
https://people.stanford.edu/pauloyer/
https://twitter.com/pauloyer

The two screenshots below appear to provide background as to how Uber management eventually (a couple of years later) chose how to address the issue. The moral of this story might be that while it is certainly important to retain passengers, it is also important to retain drivers!