Customer Quoting Behavior Analysis Report

Bingjie Zhang

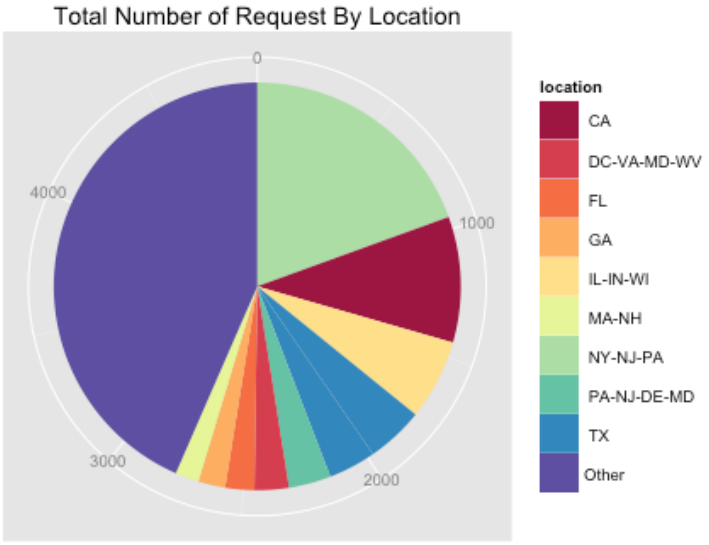
Introduction

Based on a fictional dataset of Thumbtack customer behavior, service providers' quoting behavior was analyzed to examine whether their willingness to quote had improved or declined during two months. Quoting is the activity that a service provider respond to a request sent by a consumer. By analyzing quoting behavior, we can understand more about whether requests are matched to suitable business providers.

Data Set Description

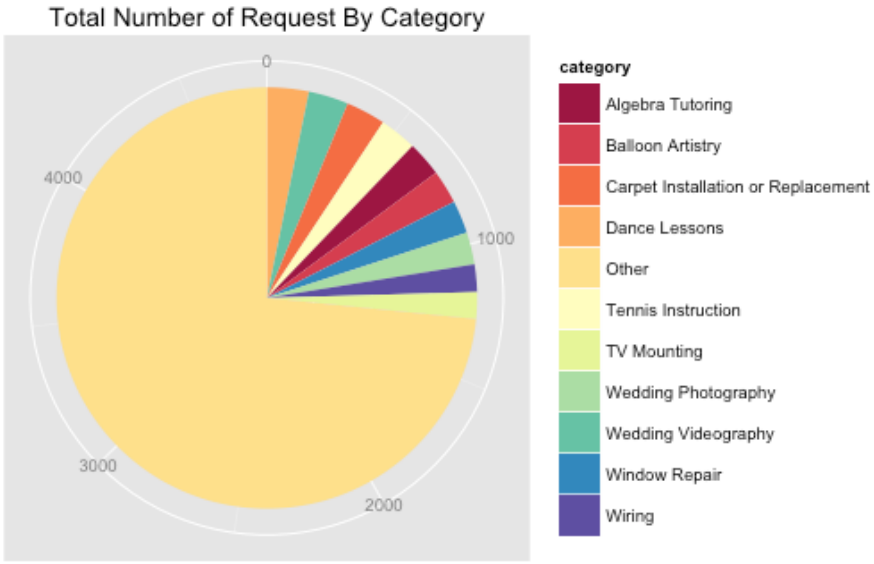
There are 10,000 observations about consumer requests and business service providers' quotes from July 1st to August 31st. Three behaviors are described in the dataset: request, invite and quote. When a consumer needs certain service, they would post a request. Then invites are sent to service providers matched by the website. Service providers would express their interest by quoting.

The dataset also contains location information of both service providers and customers and category of requests. There are 100 location areas in total. Graph 1 shows the pie chart of total number of requests by location, where top 10 locations with largest number of requests are shown. It is obvious that more than half of requests come from those 10 locations and location NY-NJ-PA takes the largest share.



Graph 1

There are 113 categories in the data set. Graph 2 shows the pie chart of top 10 categories with highest total number of requests. It shows that approximately a quarter of requests come from top 10 categories and there is no dominant category.



Graph 2

Problem Description

We are interested in analyzing whether service providers' willingness to quote had improved or declined with time. As there is no dominant request type, the effect of category will not be taken into account in the following analysis. We are interested in detecting if there exist change points that service providers behaved very differently before and after them.

To solve this problem, invite-to-quote rate is calculated for all requests. As we are interested in change during the time, day variable is created representing the day consumer posted their requests. We wanted to analyze the problem by first examining if there is any change in quoting rate by time, then examining particularly whether the mean and variance of quoting rate had changed with time.

Methods Description

As the data set contains observations from 63 days, we can subset them by time. Let Y be the set of all observations, it can be partitioned into 63 parts: $Y = \{Y_1 \cup Y_2 \cup Y_3 \cup \dots \cup Y_{63}\}$. In order to detect change point, the differences between observations happened before and after n days are analyzed. For day i, $Y(i) = \{Y_{\max(0,i-n+1)} \cup \dots \cup Y_i\}$ and $Y(i+n) = \{Y_{\min(i+1,63)} \cup \dots \cup Y_{\min(i+n,63)}\}$, we analyzed the distribution difference of Y(i) and Y(i+n) with Symmetric Pearson Divergence $PE_\alpha\big(Y(i)\big|Y(i+n)\big) + PE_\alpha(Y(i+n)|Y(i))$. Two sample t test and two sample F test are used to test if mean and variance of Y(i) and Y(i+n) are the same.

Relative unconstrained Least-Squares Importance Fitting (RuLSIF)

RuLSIF method is proposed by Liu et al.(2013) and they proved that RuLSIF is an efficient way to analyze density ratio between two samples, which can be used to calculate Pearson Divergence to measure how different are these samples distributed. α-relative PE divergence is used to analyze the distribution difference, which can be estimated with

$$\widehat{PE_\alpha}\big(Y(i)\big|Y(i+n)\big) = -\frac{\alpha}{2n}\sum_{i=1}^{n}\hat{g}(Y_i)^2 - \frac{1-\alpha}{2n}\sum_{i=1}^{n}\hat{g}(Y_{i+n})^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{g}(Y_i) - \frac{1}{2}$$

where $\hat{g}(Y_i) = \sum_{l=1}^{n} \theta_l K(Y, Y_l)$ and $K(Y, Y') = \exp\left(-\frac{\left\|Y-Y'\right\|^2}{2\sigma^2}\right)$ is the Gaussian Kernel.

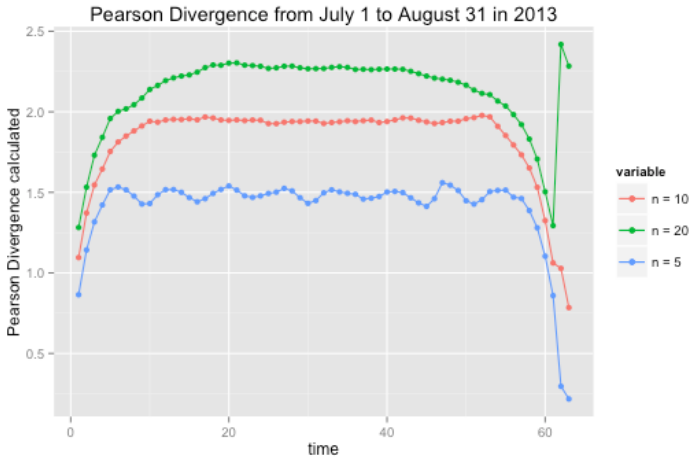
Density ratio $\hat{g}(Y_i)$ are estimated with RuLISF by learning the parameter $\vec{\theta}$ by minimizing α-squared loss between true and estimated relative ratios:

$$J(Y) = \frac{1}{2}\int p'_\alpha(Y)\big(r_\alpha(Y) - g(Y;\theta)\big)^2 dY$$

Grid search and 5-fold cross-validation are used to find best parameters to minimize J(Y). In all of the following analysis, α is set to be 0.1 because Liu et al.(2013) showed that estimated relative density ratio will always be smaller than $\frac{1}{\alpha}$. Symmetric α-relative Pearson Divergence is used because Liu et al.(2013) showed that it performed the best in detecting distribution changes. It is calculated with formula: $PE_\alpha\big(Y(i)\big|Y(i+n)\big) + PE_\alpha(Y(i+n)|Y(i))$. If the samples can be considered to be from the same distribution, Symmetric α-relative Pearson Divergence will be very close to 0.

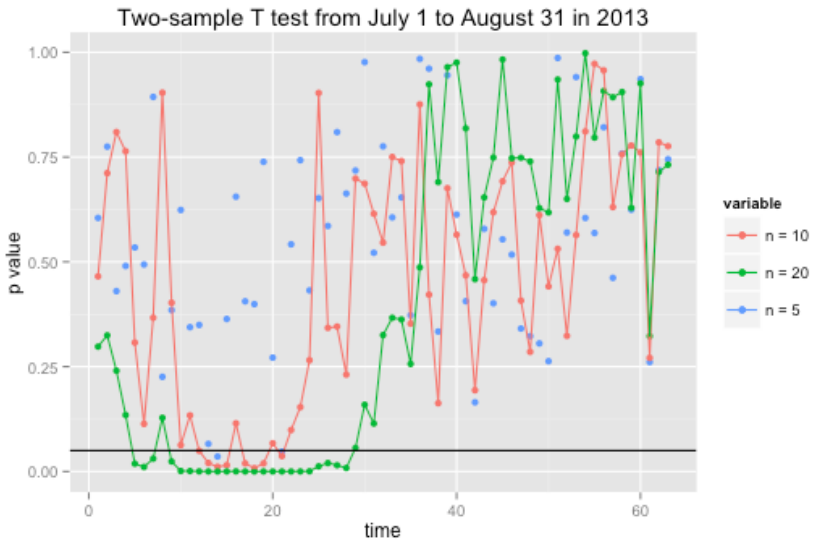
Result Summary and Analysis

Symmetric α-relative Pearson Divergence on all observations with 5 days, 10 days and 20 days before and after for each day are calculated respectively. The summarized result is shown in graph 3. It shows that the distribution of quoting rate is always changing with time and there is no peak or rapidly changing point. One potential explanation of the result is that there is no unique distribution to represent the whole quoting rate. Another way to explain the result is that quoting behavior was changing at constant rate across the time.



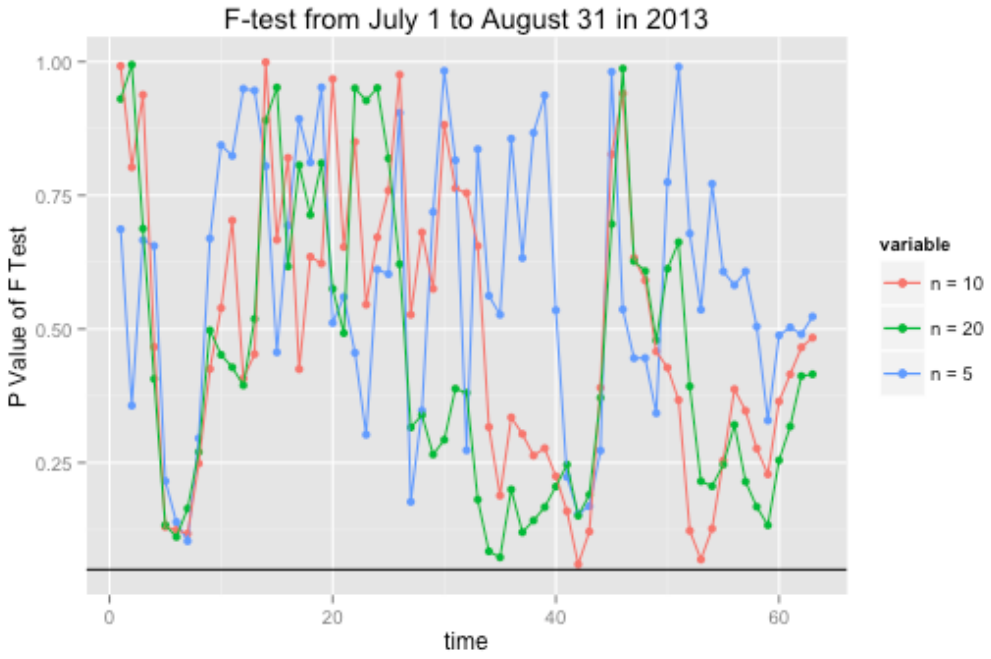
Graph 3

Two sample t-test are also performed before and after 5, 10 and 20 days before and after for each day, the summarized result is showed in graph 4. The horizontal line in the graph below is y = 0.05. From the plot below we can see that there is no statistically changes within 5 days range. However, in 10 days range and 20 days range, for certain period there are statistically significant changes. As we are interested in long-term change, July 28th can be considered as a change point because it is the last day that shows statistically significance difference before and after 20 days range.



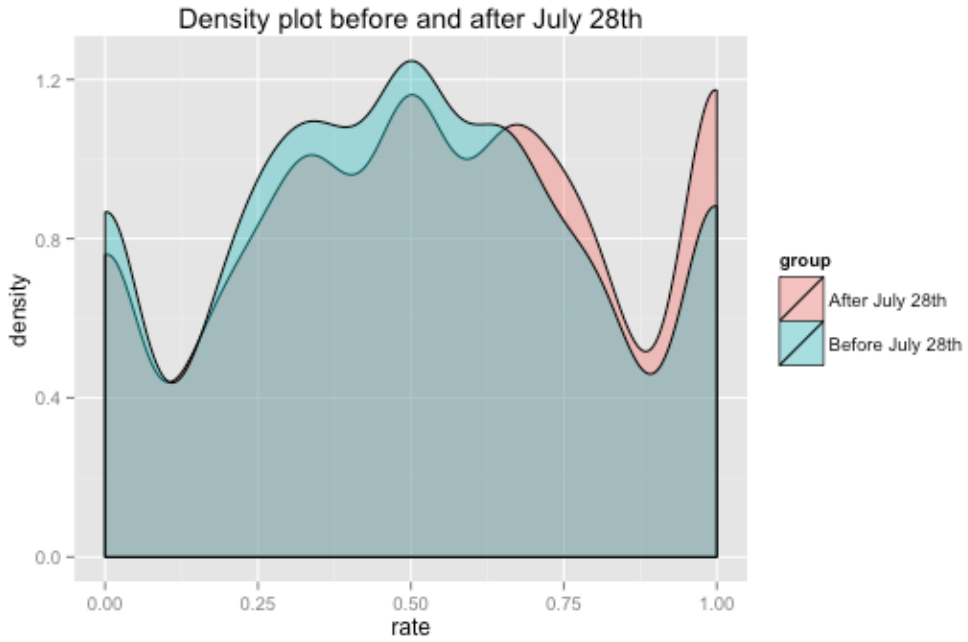
Graph 4

Two sample F-test are also performed before and after 5, 10 and 20 days before and after for each day, the summarized result is showed in graph 5. We can see that there is no statistically change in variance across the time.



Graph 5

As t-test showed that July 28th could be considered as a change point, two sample t-test is performed on observations before and after July 28th and p value is $3.001 * 10^{-5}$, which suggests that there are quite significantly mean change. The mean of observations before July 28th is 0.4992 and the mean of observations after it is 0.5362. It means that average quoting rate is gradually improving over the time. The density plots of the two observations are shown in graph 6. We can see that quoting rate after July 28th is more skewed to the right. After July 28th there quoting rates less than 0.6 are less likely to appear.



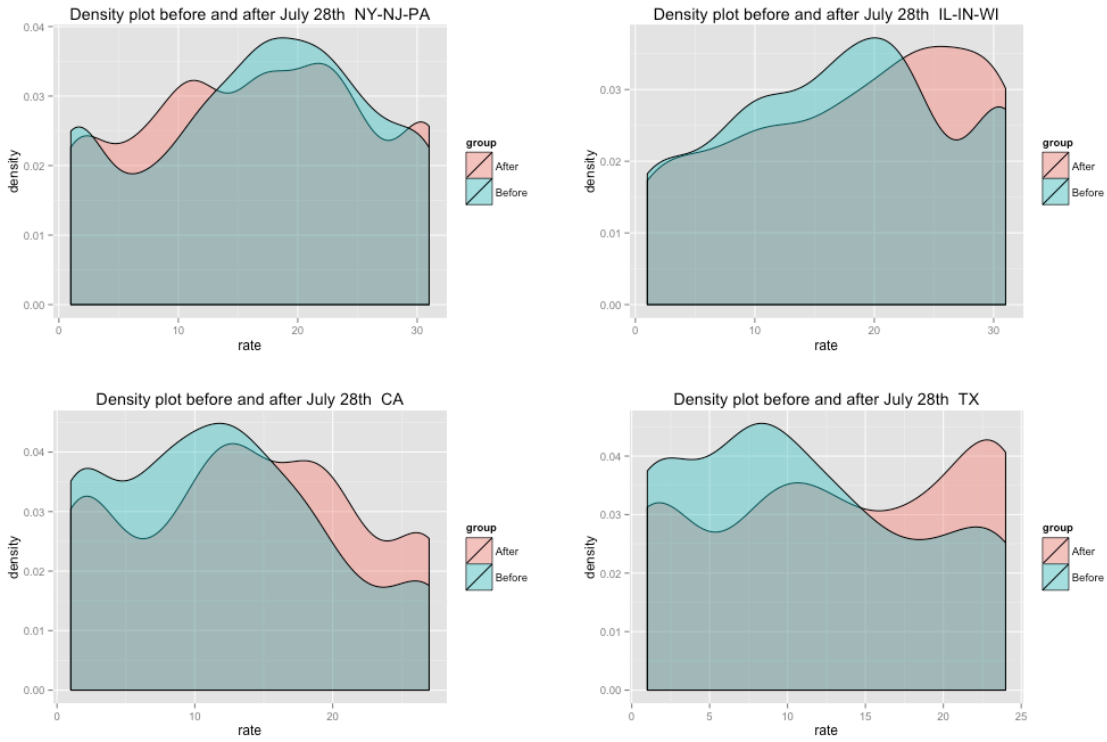
Graph 6

Symmetric α-relative Pearson Divergence of observations of top 10 locations before and after July 28th are summarized in the table below. It shows that location 1-4 showed significant change in distribution before and after July 28th while others do not seem to change very significantly.

PE divergence by Location		
ID	Location	PE divergence
1	NY-NJ-PA	1.64720589
2	CA	1.15351205
3	IL-IN-WI	0.75275767
4	TX	0.54841165
5	TX	0.25837473
6	PA-NJ-DE-MD	0.23770644
7	DC-VA-MD-WV	0.01891284
8	FL	-0.10653611

9	GA	0.02719458
10	MA-NH	-0.15700042

Density plots of locations with PE divergence larger than 0.5 are showed in graph 7. Two-sample t test are also performed to test if there is any statistically change of mean quoting rate. Only location 2 and location 4 showed statistically significant change. Average quoting rate in California (location2) and Texas (location 4) are significantly improved. Graph 7 also supported this conclusion because after July 28th, the distribution of quoting rate tilted to right to a large extent. We also notice that the distribution of quoting rate got fatter tail after July 28th. The quoting rate in IL-IN-WI also improved after July 28th, even though we do not have enough evidence to prove that the average quoting rate had improved.



Conclusion

The result above showed that probably some change happened at or before July 28th and generally speaking business provider are more likely to quote because of the change. However, when looking at the changes by location, some of the locations do not seem to be affected

dramatically. Business providers in Texas and California showed most dramatically improvement in willingness to quote. Willingness to quote of Business providers in IL-IN-WI also improved. On the other hand, willingness to quote in NY-NJ-PA seemed to decline after the change. As 19% requests come from NY-NJ-PA, certain measures should be taken to solve this problem.

Reference

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, *43*, 72-83.