The Pricing of Individual and Aggregate Consumer Information in Today’s Marketplace

Kaira Daphne Gui
kgui@seas.upenn.edu

Tania Mohd Nor
tanimn@seas.upenn.edu

Advisor
Johnathan M. Smith
Abstract

Privacy and the demand for consumer information are topics that are increasingly important with the advent of the online marketplace. The transactions and behaviors of consumers online can now easily be monitored electronically through cookies or spyware. The intersection between information security and social sciences has now been brought into focus because purely technical solutions have failed to solve the privacy issue. Consumers may be poorly motivated to protect their information if they only consider its low marginal value to themselves instead of the cumulative value it presents to companies.

Consumers regularly make trade-offs between privacy and a host of incentives such as freeware, price discounts, reward points, sample products and more. These are essentially economic transactions involving the trading of information. Companies highly value aggregate consumer data in order to execute behavioral targeting and price discrimination campaigns. The ability to segment their customers and charge personal prices enables companies to extract the maximum value along the entire frontier of the customer demand curve.

The goal of the project is to determine how the price of consumer information changes for a firm depending on the amount of data collected. The outcome of this project then gives us a starting point to think about a firm’s incentive behind collecting more data. Our project is based on a database of 100,000 customer summary records from a major wireless service provider, which includes customer demographic data, call behavior characteristics and total revenue. In the competitive world of wireless telecommunications, customer churn is a huge problem faced by service providers. And
in order to combat churn, the first most important step is to predict which customers are likely to leave, and customer information is the most vital element in doing so. Because of the importance of customer data, the trend of accumulating more information about people, and creating a detailed picture of an individual’s behavior, is well illustrated in the field of wireless telecommunications.¹
Related Work

Several surveys concerning how people value privacy have been previously conducted. One common approach is to ask participants to rate their opinion (strongly agree, agree, etc) on various privacy questions such as “Are you willing to give out your home address to this website?” This method however does not set the participant in an environment needed for an accurate response. Thus, the other approach which consists of placing participants in a hypothetical scenario through a narrative, and then asking them questions pertaining to their privacy concerns and personal information. The scenario approach is able to provide an environment for the participant, but it is text based, thus it lacks the “realness” of a visual stimulus. A method used to overcome this setback is to have participants take the survey online, and also provide them with screen shots of actual websites. Participants were then asked their willingness to provide personal information to these websites. The above mentioned surveys were used to investigate user privacy opinions, and thus did not give a quantitative value as to how concern the participants were towards giving out their personal information.

One group of researchers placed its participants into a scenario of job seekers searching for employment via different websites on the internet. These ‘job seekers’ had to make decisions that affected both their privacy as well as their hiring potential. The outcome of this experiment was that internet users were willing to pay large amounts to increase their online security.

A separate experiment had participants rank different alternatives according to their preference. These “alternatives” differed by monetary reward for their personal information, visit frequency, amount of time saved per year, and privacy concerns such as
unauthorized secondary usage of information. This experiment found that internet users were highly concerned with information privacy, and valued protective measures as well, similar to the previous group. Based on the participants’ responses, protection against errors, improper access and secondary use of personal information were valued at between $30.49 - $44.62. Besides that, they also valued monetary rewards, time saving and convenience.

An experiment conducted by Hewlett-Packard Labs studied people’s willingness to reveal personal information such as weight and age. The participant demanding the least for the information was paid the second-lowest demanded price, and in exchange, had to reveal the information to the other participants. The experiment’s financially competitive nature enabled the experimenters to price exactly the value that each individual placed on his private information. It was concluded that the less desirable a trait is to an individual, the greater the price demanded for releasing the information.

Humboldt University of Berlin conducted a shopping experiment to investigate how the request for different information units would effect consumers’ perception of private consumer information cost, or PCIC. Participants were put in an e-commerce type of environment where their data revelation would earn them the benefit of a personalized agent recommendation. Participants were also asked to judge the questions employed by an electronic agent. Results from this study show that there is a large discrepancy between online users’ expressed privacy concern and their subsequent behavior. Majority of the participants readily revealed highly personal information and were easily drawn into communicating with the virtual agent.
Firms use customer data to target ads for more efficient marketing, price discriminate against their customers, or in more general terms, cut costs and increase profits. It is for this fact that customer data is no doubt a valuable asset to a company, and the license and rights to it can be assigned, transferred, bought and sold. Thus, different firms have developed proprietary methods to price information for the purpose of marketing and corporate strategy planning. Due to this, it is easy to see why such methods are concealed from both competitors and consumers. We have found no research or articles that comprehensively addresses real price of aggregate consumer data to a firm, which is directly related to our focus on how the price of consumer information changes for a firm depending on the amount of data collected.

The experiment conducted by Hewlett-Packard Labs only focused primarily on an individual’s value for his weight and age information. All the other experiments we found only came to a general conclusion about people’s attitude and behavior towards their privacy, but not a quantitative one regarding a firm’s assessment towards the consumer’s information.
Technical Approach

The goal of our project is to derive the value of individual pieces of information in predicting churn. Firms highly value the ability to predict churn because it enables them to categorize its customers into different groups based on their probability to churn, which is in turn used to develop customer-retention strategies. We initiated the project by narrowing down the focus and examining a specific industry. The wireless telecommunications industry was chosen because of the richness of data available and its particular relevance to our research. For this project, data for 100,000 customer summary records with at least 6 months of service history from a major wireless services provider was used and analyzed. The data contained 171 potential churn predictors including customer demographic data, call behavior characteristics and total revenue.

Treenet, a data mining software was used to produce a model to accurately predict churn. Data mining is a technology often used to extract patterns within data, thus very suited to predict churn. Its function is divided into two main types; supervised and unsupervised. Supervised functions use predictors to predict a value. Regression and classification are both categorized as supervised functions. Regression is used to predict continuous values such as price, whereas classification is used to predict discrete values such as “click” or “don’t click” in targeted marketing campaigns. Unsupervised functions find the relationships and affinities in the data without a dependent variable or target. Cluster analysis is an example of unsupervised functions. Even though the question of whether a customer will leave or not is considered a classification problem, we used the “logistic binary” option in Treenet to derive the model. Logistic binary is used to predict a discrete outcome, using continuous and/or categorical predictor variables. The
probability of the response, which in this case is the probability of churn, is also produced using this option.

Even though the dataset had information of 100,000 different individuals, our version of TreeNet can only handle 8MB at one time, which is data for approximately 11,000 individuals, or 11% of the total dataset. TreeNet also automatically uses the first 8MB of data from the data table, thus a program that creates a table containing data of customers selected at random from the original dataset was written to avoid biasness.

3 separate data samples were used to generate an accurate model to predict churn. These are the randomly selected data sample, data taken directly from the beginning of the data table, as well as data taken directly from the middle portion of the data table.

<table>
<thead>
<tr>
<th>Data sample source</th>
<th>Prediction success (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Churn</td>
</tr>
<tr>
<td>Randomly Selected Data</td>
<td>95.26</td>
</tr>
<tr>
<td>Data from the beginning of the table</td>
<td>86.35</td>
</tr>
<tr>
<td>Data from the middle of the table</td>
<td>91.31</td>
</tr>
</tbody>
</table>

The randomly selected data showed the highest predictive success, and thus was used as the basis of the rest of our project. We also found via various trial and error runs that a combination of a learning rate of 0.1, influence trimming factor of 0.6, 9 nodes per tree and growing 3000 trees gives us the highest success rates. The learning rate controls the rate at which the model is updated after each stage, while the influence trimming factor is a data selection tool that selects how much TreeNet’s learning process focuses on the most important data, in order for TreeNet to avoid unreliable data. An influence trimming factor that is too low will result in bad data such as ‘outliers’ to negatively influence the predictability of the model, while an influence trimming factor that is too high will result in good data to be disregarded as well.
Figure 1: A screen shot from a run of TreeNet
We also managed to obtain the variable importance ranking for the variables of the model. This ranking reflects the contribution each predictor makes to the model’s predictive success. The most important predictor was scaled to a 100, and the other predictors’ importance were scaled relative to the most important predictor.

Figure 2: Top 10 most important variables for predicting churn
Figure 3: Impact of handset age on the dependent variable, CHURN

Figure 4: Impact of percentage change in monthly minutes used on the dependent variable, CHURN
Figure 5: Impact of percentage change in monthly revenue per customer on the dependent variable, CHURN

As an example, Verizon Wireless reported churn rates of 1.17% per month for the year of 2006.\(^8\) With 59.1 million subscribers, and their average revenue per unit (ARPU) or average revenue per month per subscriber at $50.78,\(^9\) Verizon Wireless ended year 2006 with a revenue of $38 billion.\(^{10}\) Using churn rates of 1.17% per month, implies a churn rate of 14% per year for Verizon Wireless. This means that Verizon would lose approximately 8.3 million of its customers. Assuming that Verizon’s customer retention efforts had a success rate of 15%, thus enabling them to retain 15% of the churners, Verizon would benefit from an added revenue of $63 million per year.
Interpretation of Results

Figure 6: Effect of the sample size of customer information on the predictive success of the model

Figure 7: Added annual revenue from churn prediction
Churn prediction gives us a good approximation for the value of the data given up by a single consumer. The chart above shows the revenue generated from successful attempts to retain customers who have been identified to churn in the near future versus the sample size used to identify such customers. The sample size affects the accuracy of the identification of the potential churners, and consequently, the revenue generated from retaining them. If a model only has 50% correctness, the company could only direct its marketing efforts on half of the true population of the potential churners instead of the entire group. The accuracy curves above show a sharp bend at the 300-mark in the sample size. The two lines above show the revenue generated by a 15% or 50% success rate in the companies’ attempts to retain the identified churners. These percentages reflect a range of success rates that may be realistically achieved by a company. The annual revenue per subscriber, annual churn rates and total number of subscribers, though taken from Verizon’s 2006 data, are meant to be representative of similar telecommunication companies. The slope of the linear regression through the data points between point 0 to point 300 in the sample size axis is 1.86 for the 15% line and 6.19 for the 50% line. This slope represents the additional revenue generated by a one-unit increase in the sample size. The slope from point 300 to point 11,157 in the sample size axis is zero for both lines. Thus, we can infer that the value of a single consumer’s information to the firm is approximately $1.86 to $6.19 million if the firm has collected less than 300 separate data points. However, the additional value of a consumer’s information over 300 is negligible to none. These numbers represent the lower end of the scale because they only consider the annual revenue generated by a single customer and not the entire customer’s lifetime value to the firm. However, these numbers also do not
reflect the additional marketing costs incurred by the firm in order to win back the potential churners.

The results from the churn dataset have more generalized implications for the divide between a consumer and firm’s valuation of privacy. Whereas a firm could place a price within the range of $1.86 to $6.19 million on a single consumer’s profile, this consumer may only need a $30.49 to $44.62 incentive to give it up. In exchange for this information, a firm could initiate efforts to recover soon-to-be lost profits. Also as previously posited, the value of consumer information in aggregate for the firm is not only variable but also much larger than the mere total of the potential churner’s annual revenue. The nature of a firm’s business dictates its target customers and consequently, the type of predictors, number of predictors and threshold sample size level necessary to arrive at an accurate customer identification model.

Firms who possess vast databases of consumer information such as Wal-mart, Google or other online retail sites, have the incentive to compile profiles of consumer characteristics and shopping habits to be used by their own firms, or to be farmed out to business partners, telemarketers, or direct-mail solicitors. These third-party users of data may require different queries and threshold sample sizes in order to effectively target their own consumers. By aggregating a large and broad a database, Google and similar sites may contain enough data to be relevant to as wide a variety of consumer information users as possible. Also, such firms possessing vast databases may be able to price their data offerings discriminatorily to the third-party users based on the threshold amount of data and the specialized nature of the queries required by such users.
Challenges Faced

Due to the limitations of our version of TreeNet, we were not able to analyze the entire data sample consisting of 100,000 customers. Our version, the free trial version, only allows us to analyze up to 8MB of data. We decided not to purchase the full version due to cost factors. Aside from that, the trial version did not disable any of the program’s functions. Although we were able to perform all of our stated objectives, we were unable to test the predictive capacity of the model on a larger sample size. We were also unable to verify if a larger sample size materially affected the order of variable importance.

The machine limitations also posed a significant challenge. Because of the nature of TreeNet, which requires the generation of almost 3000 trees to produce an optimal model, coupled with the size of our data and the chosen learning rate, it usually required almost an hour per run of TreeNet. The down time of each run resulted in unnecessarily slow progress. Also, in order to produce the randomized dataset, our program took almost 3 days to generate 8MB of data, which is approximately 11,000 entries.

Additionally, it would have been more beneficial to further test the model on other datasets related to online retailing. However, there were no such datasets available to us.
Conclusion

The project aimed to address the issues regarding the valuation of consumer information. This question has gained particular relevance due to the internet's increasing ability to effectively and seamlessly collate consumer information from the vast number of people who surf and conduct commercial transactions online. Firms will then be able to use such data to price discriminatorily and focus their marketing campaigns in order to increase revenues—even if it is sometimes perceived as a detriment to the consumer. Thus, it is interesting to study the divide between a consumer's valuation of his own data versus the value it brings to a firm in aggregate. Also, it is worth examining the effect of optimal sample sizes that an industry requires to extract meaningful results from the data.

In order to achieve these, we decided to use the rich amount of customer churn data that was made publicly available for the Churn Modeling competition held in 2002. With TreeNet, the winning modeling software used, we were able to identify the extra revenue generated by an additional consumer's information below an optimal sample size. Our results indicated that at a range of $1.86 to $6.19 million, the value to a firm has the potential to be many times larger than a consumer's valuation of the data. Also, the plot of the relationship between sample size and potential revenue generated by effective churn modeling indicated that at a unique point, the additional value of a one-person increase in sample size is negligible.

Due to the fact that we addressed a unique and far-reaching question that impacts many fields of study, we did not anticipate that it would be quite so difficult to gather enough relevant datasets to test our hypotheses on. Furthermore, we underestimated the cost for the full version of the software we intended to use. However, we were able to
find unconventional solutions by sourcing our information from academic institutions and making the most of the free trial version of the software. Future researchers who wish to build upon our study may want to explore the optimal sample size levels and the price of aggregate consumer information for different industries and draw conclusions about these arrays of values. Furthermore, they may wish to determine the range of prices placed on a variety of individual variables that are relevant to a variety of industries such as location, occupation, mean income or credit score. Knowing this would enable future researchers to comment on the value that individuals and consumers should place on single queries.
References


3 Ibid.

4 Ibid.


9 Ibid.
