# Predicting Customer Behavior Using Data – Churn Analytics in Telecom

Tzvi Aviv, PhD, MBA

## Introduction

In antiquity, alchemists worked tirelessly to turn lead into noble gold, as a by-product the sciences of chemistry and physics were created. In our post-modern era, 'data scientists' are working to generate valuable knowledge from mountains of data using statistical methods. Here, I will use data analytics to predict customer behaviour and gain business insight for value capturing by telecom companies. The rapid adoption of mobile technology by consumers created a saturated wireless market in Canada, dominated by three major networks (Rogers, Telus and Bell, see **Figure 1**). Annual revenues in the Canadian wireless sector are 22B dollars, with a lucrative 37% EBTIDA margin[1]. Currently, there are twenty-nine million wireless subscribers in Canada, corresponding to 80% of the Canadian population. While phone subscription grew slowly over the last eight quarters, it is expected that subscription will soon reach saturation. The technological shift from plain cell-phones to smartphones benefited the bottom lines of telecom companies, as smartphone customers pay higher monthly bills in data usage fees. However, the shift to smartphone is nearly complete and about 80% of wireless subscribers are now using smartphones. As long as the costs of maintaining current customers are lower than the costs for acquiring new customers, telecom companies are likely to invest in monitoring churn behaviour of customers, with an aim to reduce churn and increase revenues. Voluntary churn occur when a customer is leaving one company and taking his business to a competitor company. For example, Rogers Wireless reports an average monthly churn of 1.3%, the highest churn rate among the big three telecom companies in Canada. This monthly rate may seem low, but it adds up to an annual churn rate of 15%, while total annual growth in subscribers in Rogers is 4.5%. Reducing churn rate by a third from 15% to 10% could double the growth in customer base for Rogers. The average annual revenue per customer in the Canadian Telecom industry is about $758, suggesting that a 5% growth in customer base may be translated into about $374 million in additional revenues for Rogers.

[1] http://www.crtc.gc.ca

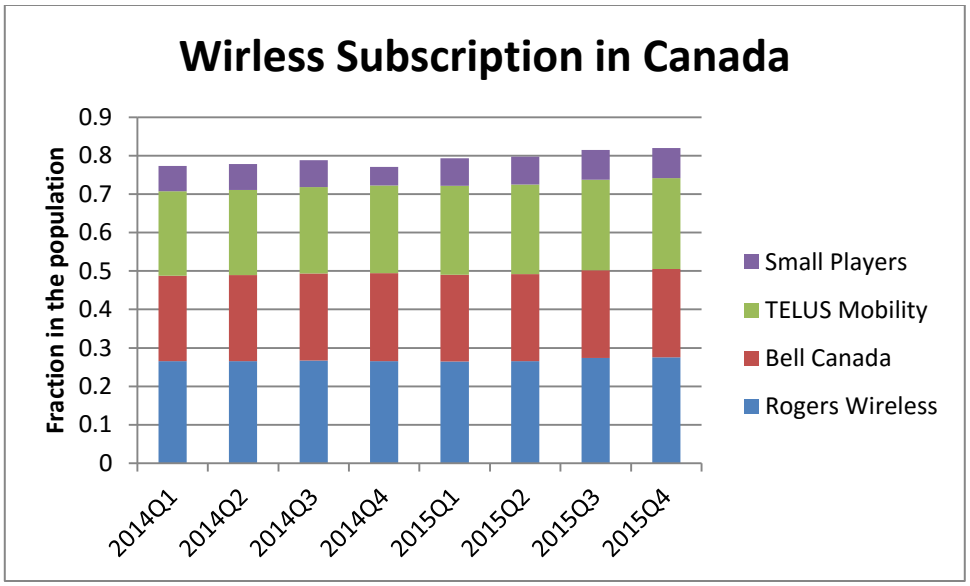

Figure 1: Wireless subscription in Canada as population fraction (Data Sources: www.cwta.ca and www.statcan.gc.ca).

To dissect churn behaviour of wireless customers, demographic information, usage data, and financial information are combined to create large datasets. Analyzing these datasets yield important business insights including possible reasons for churning. Importantly, if churn can be predicted, customers that are about to churn can be identified prior to churning and targeted directly with specific marketing promotions in an attempt to prevent churn.

**Data and Methodology**

I will demonstrate churn analytics using a publicly available dataset acquired by a telecom company in the US [2]. This dataset contains 21 variable collected from 3,333 customers, including 483 customers labelled as churners (churn rate of 15%). Surveying the churn literature reveals that the most robust methods for creating churn models are logistic regression models and regression decision trees. To avoid 'overfitting' of the data, the dataset will be randomly divided to a training set and a validation test. The training set will be used to develop the statistical model, and the validation set will be used to test the validity of the model to accurately 'predict' churning behaviour. In reality, churn models need to be validated on future data, to establish the predictive power of the model. The quality of the churn model can be assessed by multiple factors, here I will use a 'lift value', or the ratio of the percentage of actual churners in a list of predicted churners over the general churn ratio, as a quick way to illustrate the predictive value of the model. A 'lift value' above one is generated by

[2] https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383

model better than a random list of customers, while lift values above three are often quoted for good predictive models.

Analytic framework:
1. Download data and import into SAS and SPSS
2. Examine data, clean missing values, merge datasets if necessary (here the data was ready for analysis).
3. Partition data into training set (80%) and validation set (20%)
4. Explore variables and identify prediction factors
5. Use Logit function in SAS or Decision trees in SPSS to create a predictive models
6. Assess the predictive models using the validation set, calculate lift value

**Results**

Exploring correlations among of the variables in the dataset reveal six variables with statistically significant association with churn behaviour, these variables will be utilized to generate a predictive model. Inspecting the categorical variables using SAS reveals that customers with an international plan are fourfold more likely to churn (**Figure 2**, $\chi^2 < 0.001$). The company should inspect its international plan and how it fits the needs of customers. Conversely, customers with a voice mail plan are twofold less likely to churn than customers without a voicemail plan (**Figure 3**, $\chi^2 < 0.001$). Inspecting the distribution of numerical variables among churning and non-churning customers also reveals interesting insights. While only 5% of non-churners called for service more than three times, 27% of churners had done so (**Figure 4**). The company should monitor the service quality in its call centers in an attempt to resolve technical problems in less than three calls. In addition, customers calling more than three times should be flagged for marketing promotion. It is also evident that many churning customers are using more day-time minutes, and as a result are paying higher bills (**Figure 5** and **Figure 6**). These finding indicate that many churning customers are ‘high usage’ customers, these are more valuable customers for the company due to the higher fee they pay. Unfortunately, high charges may be pushing these customers to find a cheaper plan or a cheaper provider. Targeting these high usage customers with price promotion may provide an incentive for them to stay as customers.

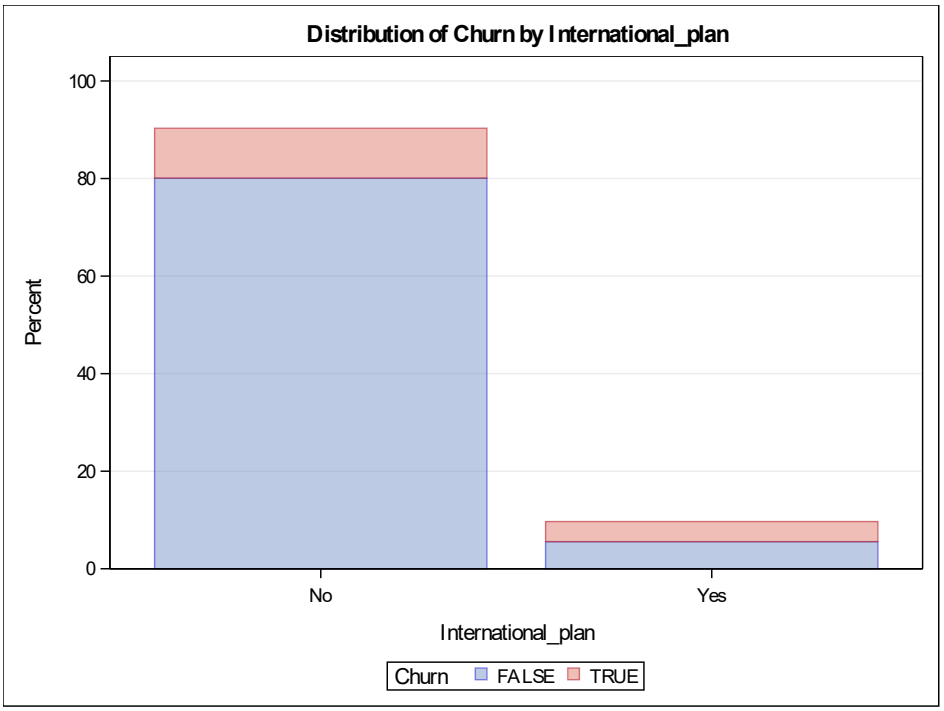

Figure 2: Distribution of Churn by International plan.

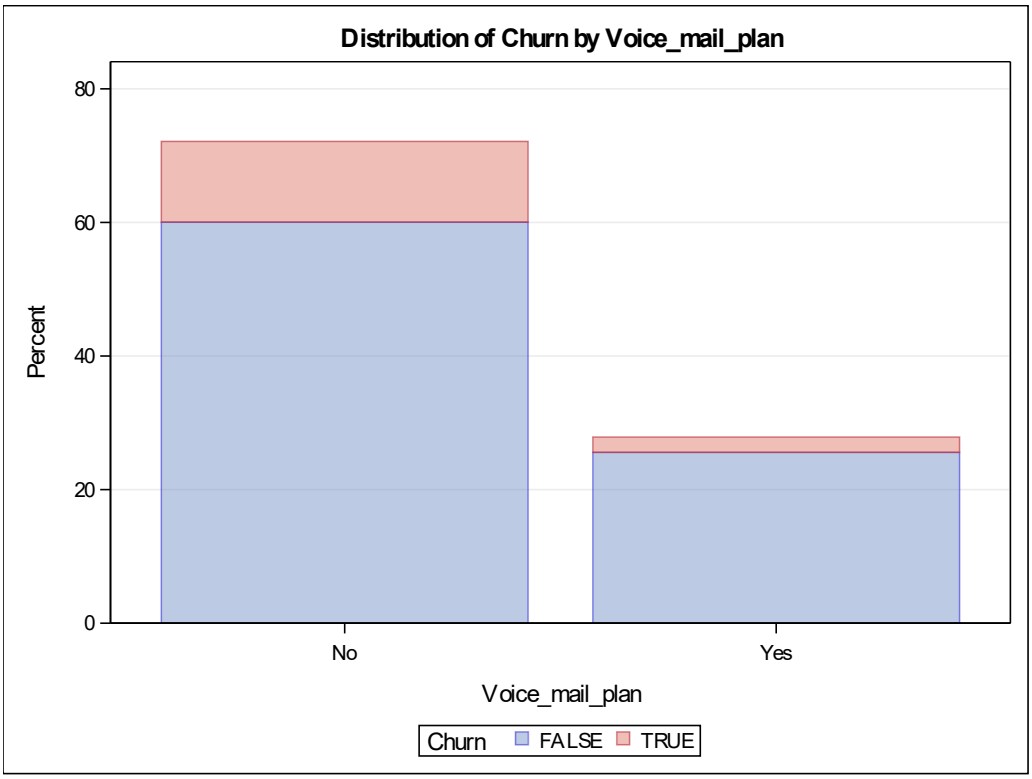

Figure 3: Distribution of Churn by Voice mail plan.

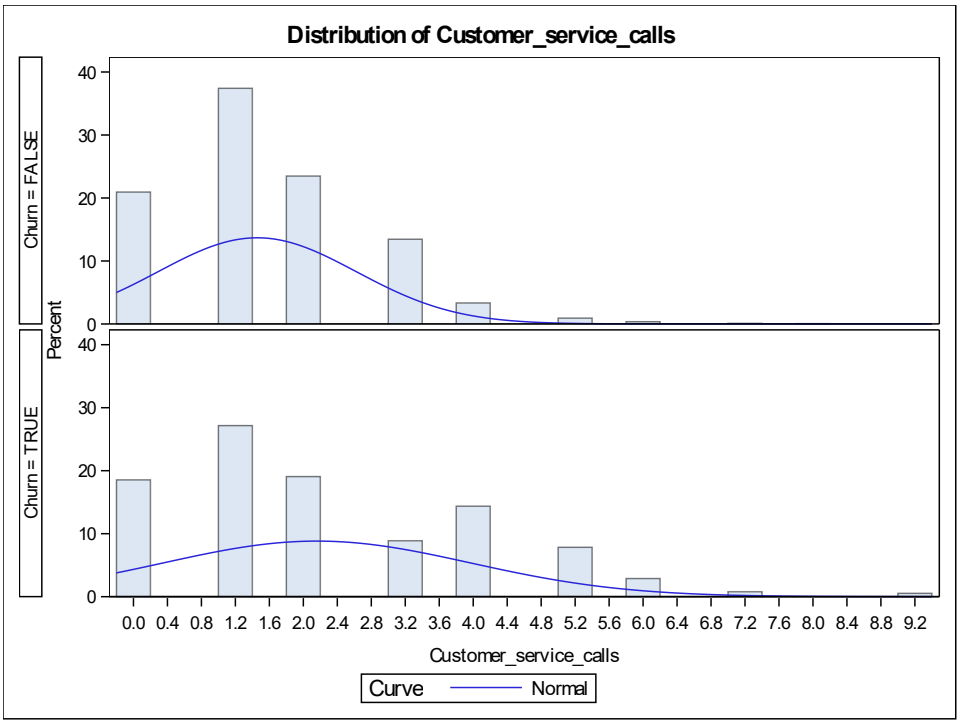


Figure 4: Distribution of customer service call among churning and non-churning customers. Notice the skew to the right in the lower graph.

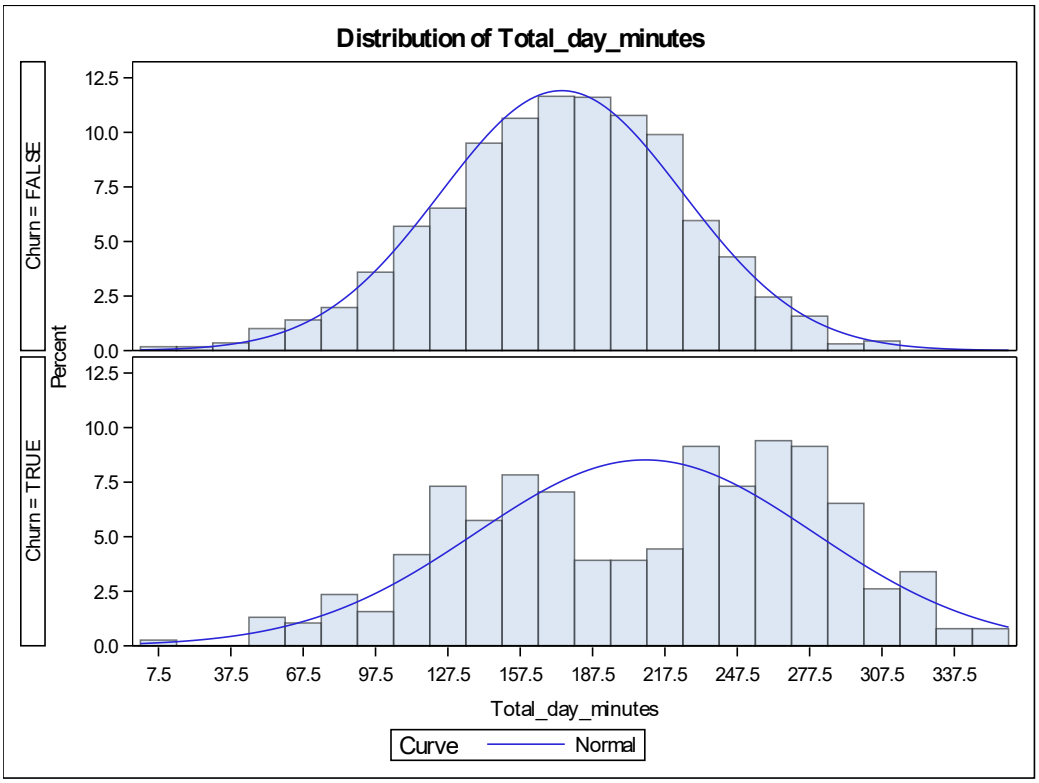


Figure 5: Distribution of day time minutes among churning and non-churning customers, notice the rightward skew among churners indicating high day time usage.

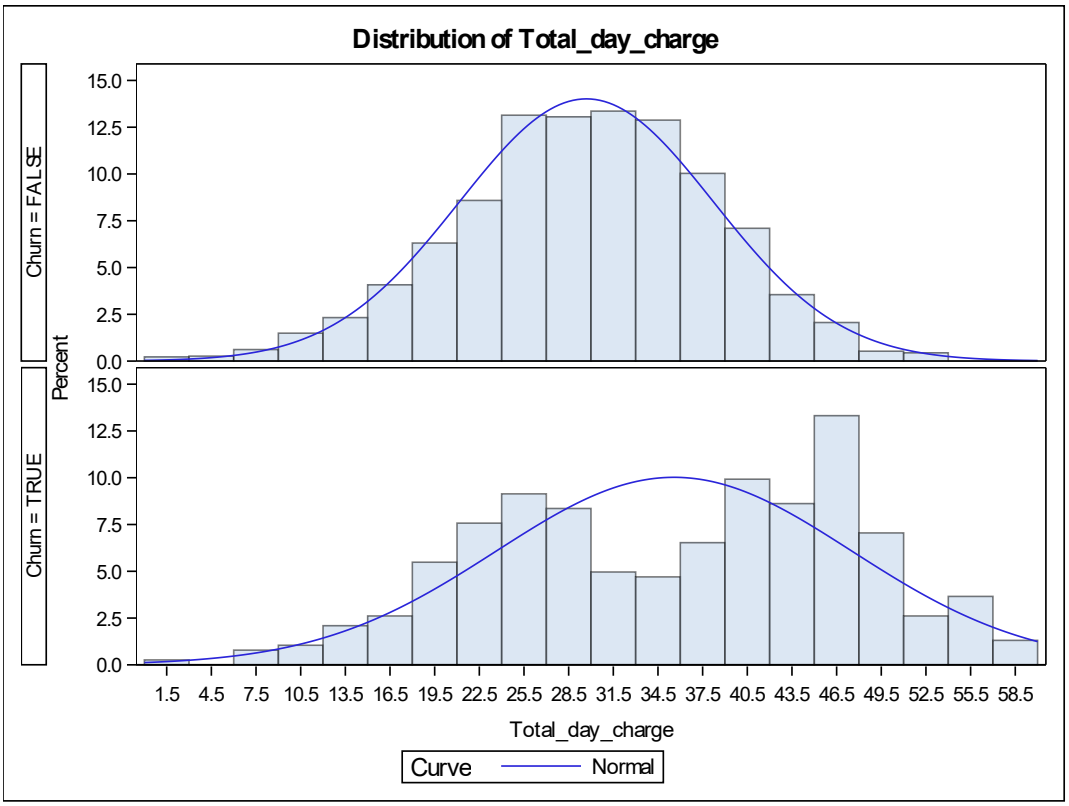


Figure 6: Distribution of day charges among churning and non-churning customers. Notice the rightward skew in the lower panel, indicating higher fees payed by many churners.

**Model 1: Logistic Regression**

Combing the significant predictors of customer churning into a regression model that includes the following variables:

1. International plan – as a categorical variable
2. Voice mail plan – as a categorical variable
3. The amount of monthly day-charge or day-minutes (these are correlated to each other, and thus redundant)
4. The number of service calls

Using these variables, a logistic regression model with a concordance value of 0.81 was computed in SAS using the training set. Moreover, when this model was tested on the validation set, a similar concordance value of 0.8 was computed, indicating no overfitting of the model. The regression model provides the probability of a customer to churn based on the indicated variables. When a probability of 0.4 is used as a cut-off to indicate churners, the model predicted 56 of 533 customers in the validation set as churners. Comparing the predictions of this model to the actual status of the customer

revealed that 18 of the 56 customers are truly churners (a modest lift value of two, **Table 1**).

| Table 1: Churn Prediction by Model 1 | | | |
|---|---|---|---|
| | Churn | | |
| Churnpred | FALSE | TRUE | Total |
| 0 | 407 | 70 | 477 |
| | 85.32 | 14.68 | |
| 1 | 38 | 18 | 56 |
| | 67.86 | 32.14 | |
| Total | 445 | 88 | 533 |
| | 83.49 | 16.51 | 100 |
| lift=32.14/16.51=2 | | | |

## Model 2: Decision Tree Model

A five level decision tree model of churning was created in SPSS using the variables of day charge, customer service call, international plan, and voice mail plan (in this order of importance, **Figure 7**). The resulting model labelled 254 cases out of 2,666 as likely churners. Out of these 254 cases, 174 are truly churners (68.5% or a lift of 4.5, **Table 2**).

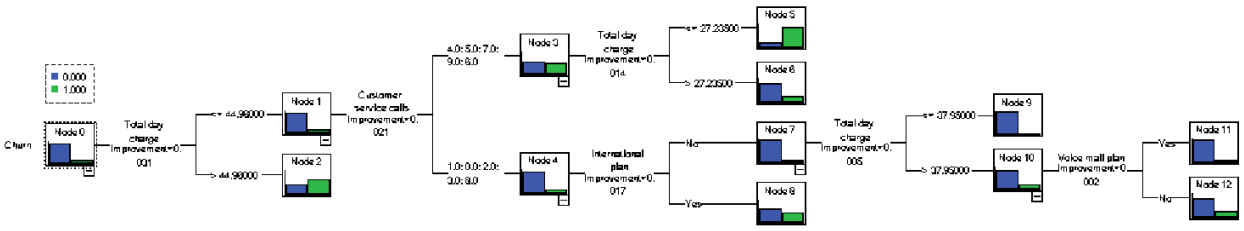


Figure 7: A diagram of the CRT model created for churning (1=Churn).

**Table 2: Model 2 Assessment**

| | Observed | | |
|---|---|---|---|
| Predicted | FALSE | TRUE | Overall Percentage |
| 0 | 2203 | 209 | 90.5% |
| 1 | 80 | 174 | 9.5% |
| Total | 2283 | 383 | 89.2% |

## Conclusions

Annual churn rates in the wireless industry are currently about 10%, while customer base grow more slowly at about 5% annually. Reducing customer churn can have a large impact on customer base. In addition, higher costs associated with acquisition of new customers highlight the need of telecom companies to reduce churn rate in order to decrease costs and increase revenues. Data analytics can yield important insights about customer behaviour and may contribute to churn reduction. A data set of 3,333 US wireless customers, including 483 churning customers, was dissected here to reveal:

1. High churn rate among customers with international plans.
2. High churn rate among intense day time users.

It is likely that higher bills are driving some of these customers to look for cheaper options.

3. Customers with four or more service calls are more likely to leave the company. Companies should improve their service call centers to resolve customer issues in fewer than three calls.

Importantly, these findings can be integrated into a predictive model to identify customers most likely to churn, these 'on the fence' customers should be targeted for an upgrade of monthly plan or other incentives to prevent churning. The most promising churn model to fit this dataset is a CRT decision tree model – yielding a 68% 'hit rate' in the predicted churning customers list, capturing almost half of all churners. Using this model may reduce churn rate by up to 50%, leading telecom revenues to increase by millions of dollars.

## Limitations

1. The model was derived from wireless customer data in the US market, Canadian customer may have different behaviours. A better Canadian model should be derived from Canadian data and tailored to the Canadian market.
2. Customer behaviour is dynamic, and may change over time, the model should be validated periodically to assess changes in customer behaviour.
3. Churn insights are based on correlating behaviours, correlations does not necessarily indicate causative relations. Interventions in customer behaviour should be empirically tested.

## Appendix 1: Description Statistics of Numerical Variables in the Dataset

| Variable | N | N Miss | Minimum | Mean | Median | Maximum | Std Dev |
|---|---|---|---|---|---|---|---|
| Account_length | 2666 | 1 | 1.0000000 | 100.9433608 | 101.0000000 | 232.0000000 | 39.5737820 |
| Area_code | 2666 | 1 | 408.0000000 | 437.9084771 | 415.0000000 | 510.0000000 | 42.7443189 |
| Number_vmail_messages | 2666 | 1 | 0 | 8.1807952 | 0 | 51.0000000 | 13.7653732 |
| Total_day_minutes | 2666 | 1 | 0 | 179.6481245 | 179.2000000 | 350.8000000 | 54.9171794 |
| Total_day_calls | 2666 | 1 | 0 | 100.4864966 | 101.0000000 | 165.0000000 | 20.1296363 |
| Total_day_charge | 2666 | 1 | 0 | 30.5407577 | 30.4600000 | 59.6400000 | 9.3359410 |
| Total_eve_minutes | 2666 | 1 | 31.2000000 | 201.2355589 | 201.7000000 | 361.8000000 | 50.7638785 |
| Total_eve_calls | 2666 | 1 | 12.0000000 | 100.3270818 | 101.0000000 | 170.0000000 | 19.9737988 |
| Total_eve_charge | 2666 | 1 | 2.6500000 | 17.1052438 | 17.1450000 | 30.7500000 | 4.3149279 |
| Total_night_minutes | 2666 | 1 | 23.2000000 | 200.9040885 | 200.6500000 | 395.0000000 | 50.4412476 |
| Total_night_calls | 2666 | 1 | 33.0000000 | 99.9804951 | 100.0000000 | 175.0000000 | 19.6289309 |
| Total_night_charge | 2666 | 1 | 1.0400000 | 9.0407464 | 9.0300000 | 17.7700000 | 2.2699197 |
| Total_intl_minutes | 2666 | 1 | 0 | 10.2468867 | 10.3000000 | 20.0000000 | 2.7756796 |
| Total_intl_calls | 2666 | 1 | 0 | 4.4756189 | 4.0000000 | 20.0000000 | 2.4795536 |
| Total_intl_charge | 2666 | 1 | 0 | 2.7671680 | 2.7800000 | 5.4000000 | 0.7493404 |
| Customer_service_calls | 2666 | 1 | 0 | 1.5577644 | 1.0000000 | 9.0000000 | 1.3022717 |