Modelling Customer Lifetime Value for a Multinational Telecoms Company

Briain O'Donnell

The thesis is submitted to University College Dublin in part fulfillment of the requirements for the degree of MSc Statistics

UCD School of Mathematics & Statistics

Supervisor: Dr. Riccardo Rastelli

Submission: 08/2019

Student Number: 13352026

Table of Contents

List of Figures

List of Tables

1 Abstract

The primary goal of this master's dissertation is to develop a framework for modelling the Customer Lifetime Value of the customers of a multinational Telecoms company. As such, the modelling framework must be applicable across geographies. This research has been carried out in coordination with a data analytics company as a project for one of their clients.

Beginning with an in-depth data exploration exercise, various modelling frameworks are explored. The Pareto/NBD family of models was deemed the most suitable and proves to be effective in predicting the number of transactions customers will make in the future. The aim of these models is to analyze past transaction data to predict whether the customer will be alive at any time in the future and ultimately predict the number of transactions they will make. The Pareto/NBD model was applied successfully, but it was noticed that some of the data seemed to violate the NBD assumption of the model. As such, the Pareto/GGG (Platzer & Reutterer 2016) was applied to the customer cohort exhibiting patterns in violation of this assumption. The Pareto/GGG clearly outperforms the former model for this customer cohort.

This comparative advantage is leveraged as the dataset is divided into two cohorts based on customer regularity. When each individual models' performance is compared to the performance of the combination of both models, we see more accurate predictions on an aggregated and disaggregated level for the combination of both models.

Once the most suitable model is chosen, it is used to analyze the customer data and make predictions about future customer activity and future transactions. These predictions are invaluable to a large organization that needs to plan ahead to understand whether they have the necessary operational resources and to understand which customers are driving the most value to the company.

2 Introduction

Competition in the telecoms industry has intensified over the last few decades with the emergence of technologies such as Skype, WhatsApp etc. Telecom service providers need to focus on building customer relationships with the right customers, in other words, the customers who will generate positive net returns over their lifetime. Kotler and Armstrong (1996) define Customer Lifetime Value (CLV) as: "The returns over time from a person, household or company, that exceed the cost of acquiring the customer and delivering the product".

In the Telecoms Industry, retaining high value customers is paramount. Companies must have a strong understanding of the customer behaviors that are indicative of future positive net present value. This can be determined by explicitly modelling CLV. These high value customer archetypes become the target of lead generation and the focus of the customer retention efforts. Accurately modelling CLV is a complex process, however, and requires a different approach based on customer types and the unique situation. For this dissertation, I am analyzing the telecoms subscription data for prepaid (non-contractual) customers. The goal of this dissertation is to develop a suitable framework to accurately predict CLV for these prepaid Telecoms customers.

In a CLV modelling framework, a customer's status can be "alive" or "dead". When a contractual customer cancels their subscription plan, it is self-evident that they are "dead". In a non-contractual situation, the customers subscription runs out and it is not clear whether they are actually "dead" or not. The key idea is to predict whether the customer will be alive at any stage in the future, to predict the number of purchases they will make and to evaluate the net present value of these purchases. There are serious challenges to solving this problem, however. In the prepaid customer setting, the churn event is not directly observable, but needs to be inferred indirectly based on observed periods of inactivity.

We also need to account for the fact that every customer behaves differently. As such, we will need statistical methods to evaluate cohort level patterns as priors for estimating customer level quantities.

To accurately model customers behaviour, we must also attempt to understand the timing patterns of customer purchasing to understand how they consume mobile credit. Customers purchase mobile credit and use it over time. As such, even though the customer will consume credit in a random manner, perhaps they are likely to purchase credit in more regular intervals.

Once we have modelled customer behaviour, we will then attempt to predict what will happen in the future. From a business perspective, we are trying to answer the following questions as listed by Michael Platzer (2016):

A. How many customers will the company still have after x number of months?
B. How many customers will be active (regularly purchasing) after x number of months?
C. Identify which customers are considered to have churned
D. How many transactions should be expected during the next x months?
E. Identify which customers will provide the most value to the company over the next x months

Answering A and B will allow the company to understand the composition of their customer base in the future. This would inform decisions such as whether they should increase marketing spend to acquire more customers etc. Answering C would enable them to target customers who have churned with different marketing campaigns to try to re-acquire them.

Question D allows the company to predict future revenue and understand whether they will be able to meet their financial obligations etc. Question E allows them to understand who their most valuable customers are. These are the customers that the company will want to retain and therefore should be treated with preference.

3 Research & Exploratory Analysis

3.1 Initial Research

As discussed, the goal of this dissertation is to accurately predict customer lifetime value (CLV) for Telecoms customers. I began my research by studying the available literature and previously applied methods.

Tirenni (2005) describes the fundamental way to model Customer Lifetime Value (CLV) as the sum of all excess cash flows during the time period discounted to a certain factor d. For a customer i, this is expressed by the following:

$$CLV_j = \sum_{t=0}^{T} \frac{(r_{tj} - c_{tj})}{(1+d)^t} \tag{1}$$

Here we assume that the cash flows are known for each time period, which is a problematic assumption to make in real life problems. This deterministic model would give an easy and accurate measure of CLV when all cash flows are known. Unfortunately, this is rarely the case. We need to develop a model that accounts for the stochastic nature of CLV.

Floral and Friberg (2003) employ a Markov Chain Model which considers each state in the chain to represent a person being a customer for one month, with an infinite number of states. The transition probability to move from one state to the next is equivalent to a customer staying with the operator until the following month. However, the authors note that this approach has its weaknesses as the estimates become very sensitive to the input data which is a major flaw in any CLV modelling framework.

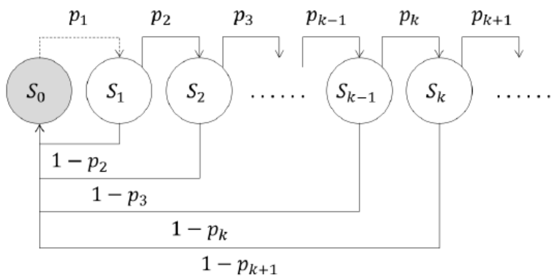


Figure 1: Markov Chain Model

Continuing my research into the subject, I came across a modelling framework first introduced in a seminal article written by Ehrenberg (1959) who proposed using the Negative Binomial Distribution (NBD) to model repeating consumer buying patterns. There has been a lot of research into this subject following this article, which has resulted in many different modelling frameworks including the Pareto/NBD model and different variants of the same. These are effectively extensions of the NBD model but make additional assumptions regarding the dropout process and heterogeneity across the customer base.

This family of probability distributions is best suited for our goal. The modelling framework I have employed makes use of these models and the remainder of this dissertation will explore their applicability and effectiveness in modelling the CLV of Telecom customers.

3.2 Planning and Execution

As mentioned, this dissertation is being undertaken in cooperation with a data analytics company to develop a framework to model the CLV of the customers of their largest clients, operating in the Telecoms industry. Once I had completed my research about the different modelling frameworks, I presented my findings to the data analytics company to align on the modelling approach and to plan the following steps. Once we signed off on the modelling framework, I got to work extracting the necessary data from their database.

We decided to use data from the largest market served by this Telecoms company. I studied the database schema, in the form of a relational database with hundreds of tables of information relating to the customers themselves and customer activity.

Fader, Hardie and Lee (2005) simulated a period of 104 weeks to fit their model. I chose a similar timeline of 104 weeks from 1st-Feb-2017 to 1st-Feb-2019 which was accessible in the data analytics companies database. The cohort of customers chosen each made their initial transaction within the first three months of this timeframe. I further filtered this customer set to customers who made at least 11 transactions during the overall timeframe, which was a necessary filter for part of the analysis to come.

To extract the data, I built a Python script to connect to the database and query the tables using SQL. This involved looping though 24 months of data tables to query each for the transaction data relating to the customer cohort.

3.3 Exploratory Data Analysis

This section will provide an in-depth descriptive analysis of the dataset. We will look to explore characteristics of the data to provide potentially valuable insights to interpret our modelling results in later sections.

3.3.1 Key Summary

The following is an overview of the primary descriptive metrics of the dataset

No. Customers	5,335
No. Transactions	324,976
Date Range	1st-Feb-2017 – 1st-Feb-2019

Table 1: Overall Data Summary

Now we provide a statistical summary of some metrics grouped at the customer level.

	No. Transactions per Cust	Transaction Amt per Cust	ITT per Cust
Average	44	144.43	4 days
SD	51.4	167.72	4.77 days
Range	[12, 405]	[52.81, 2340]	[1 day, 46 days]

Table 2: Data Summary per Customer

We have a large, heterogeneous cohort of customers in our dataset. There is considerable heterogeneity among customers regarding their purchase frequency, transaction amount and in their inter-transaction times (ITT).

Note that we filtered the dataset for customers with more than 11 transactions, which was necessary for some later analysis. There is a large distribution of transaction frequency however, with some customers purchasing over 400 times during the 24-month timeframe. Conversely, we did have some customers with a mere 12 transactions leading to a median value of 44 transactions per customer.

Similarly, we see a large distribution in the median transaction amount per customer. Some customers spend a lot per transaction, 2340 (€15.52), versus some customers who spend relatively less around 52 (€0.34).

The distribution in ITT for customers is interesting. Here we take the median transactions per customer, so the result won't be skewed by outliers. We see that the median average ITT is very low, 4 days with a standard deviation of 4.77. This indicates that, while customers are alive, they tend to transact very frequently.

The timeframe that we are considering is short relative to the average lifespan of a mobile user. As such, we must note that we are not observing the entirety of the customers relationship with the mobile service provider by considering the two-year timeframe. The total observation period is 730 days, and we filtered our dataset so that each customer's first transaction was within the first three months of the timeframe. Therefore, the maximum lifetime for a customer is ~640 days. In figure 2 below we can see the distribution of "observed customer lifetime", as calculated by the difference between the customers maximum and minimum transaction date. We would expect that most customers would still be alive at the end of the window.

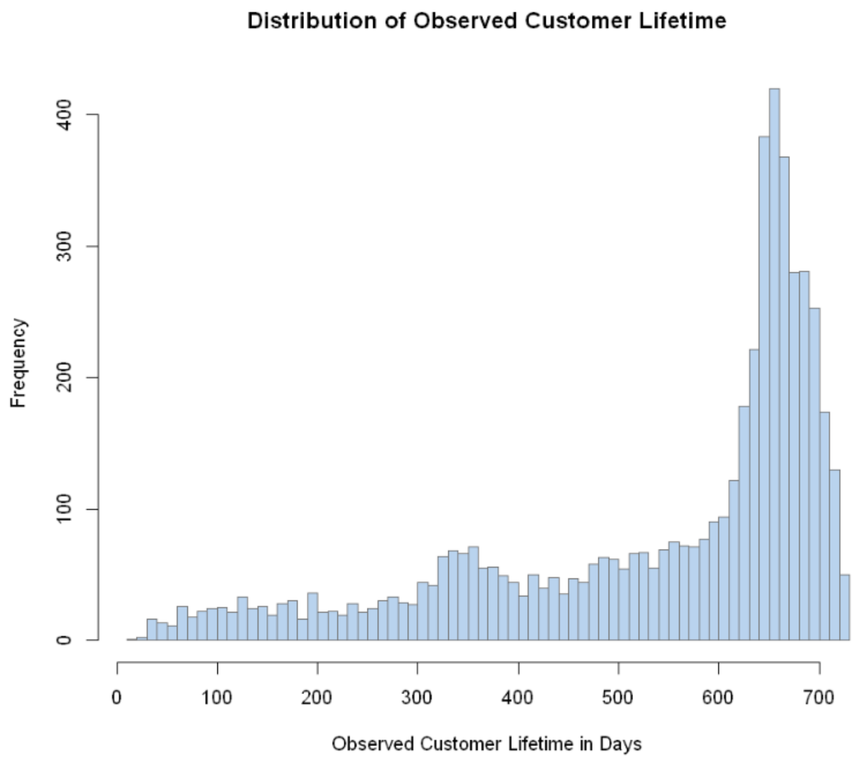


Figure 2: Distribution of Observed Customer Lifetime

Since we have data on the exact date of each transaction, we can inspect the exact transaction timing patterns of each customer. This is valuable information and should be analyzed to look for trends in customer purchasing behaviour. In figure 3 below we can visualize the transaction patterns for 10 randomly selected customers. Immediately, we see varying patterns in customer purchasing behaviour. Some customers seem to purchase in recurrent intervals, before suddenly stopping. Some customers are more sporadic and exhibit more random purchasing patterns. We will consider these patterns in our analysis, particularly the apparent defection of highly regular customers once they stop purchasing.

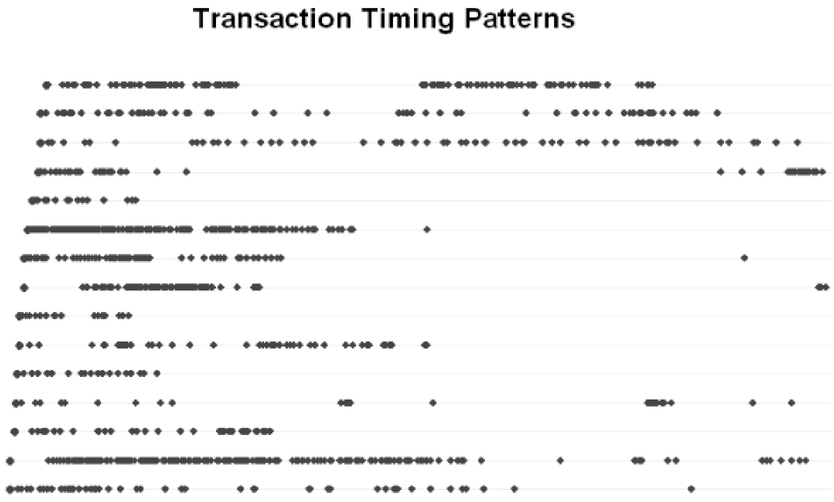


Figure 3: Customer Transaction Timing Patterns

3.3.2 Distribution of Individual Customer Behaviour

Figure 5 is a graphical representation of the distribution of the count of transactions per customer during the 24-month timeframe. It reflects the results we saw in the summary table, which showed a median of 44 transactions with a standard deviation of 51.4. The chart is heavily right skewed, reflecting the reality that most customers would be classed as “inactive” users. In figure 4, we see a breakdown of the number of customers that we would class as very active, active, inactive and very inactive.

Of course, low numbers of transactions can indicate two scenarios. 1) Customers just rarely purchase plans with their mobile credit, or 2) they drop out and their lifespan is cut short. This is one of the main challenges in modelling customers behaviour. We need to determine whether the customer is still alive at any point in time, to understand which of the above scenarios is true.

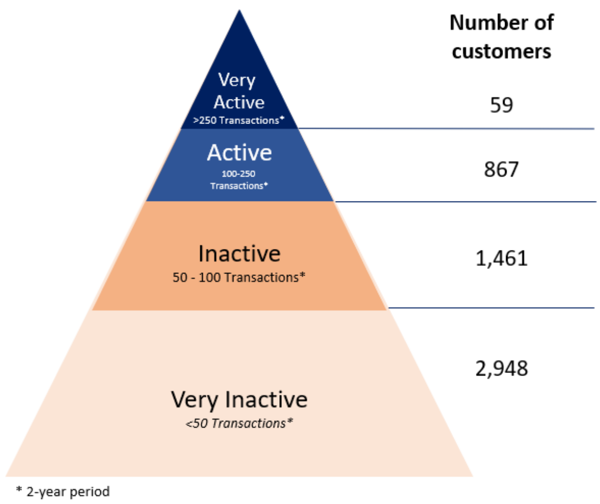


Figure 4: No. Customers in each Activity Level Bucket

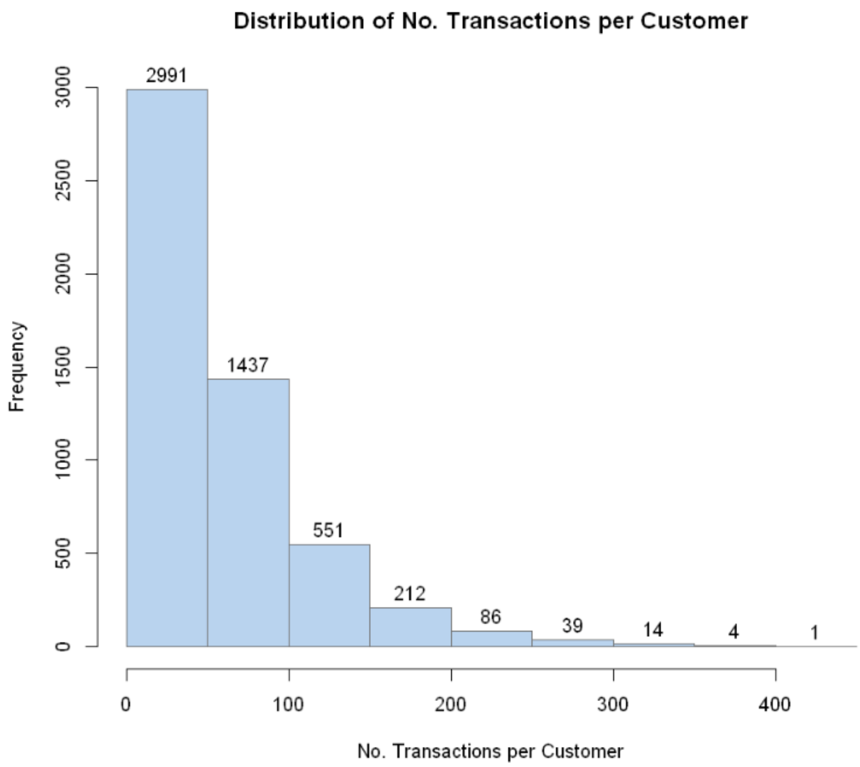


Figure 5: Distribution of No. Transactions per Customer

Next we look at the distribution of total sales per customer in figure 6. As expected, this chart is very similar in shape to figure 5, since total sales correlates strongly with total transactions.

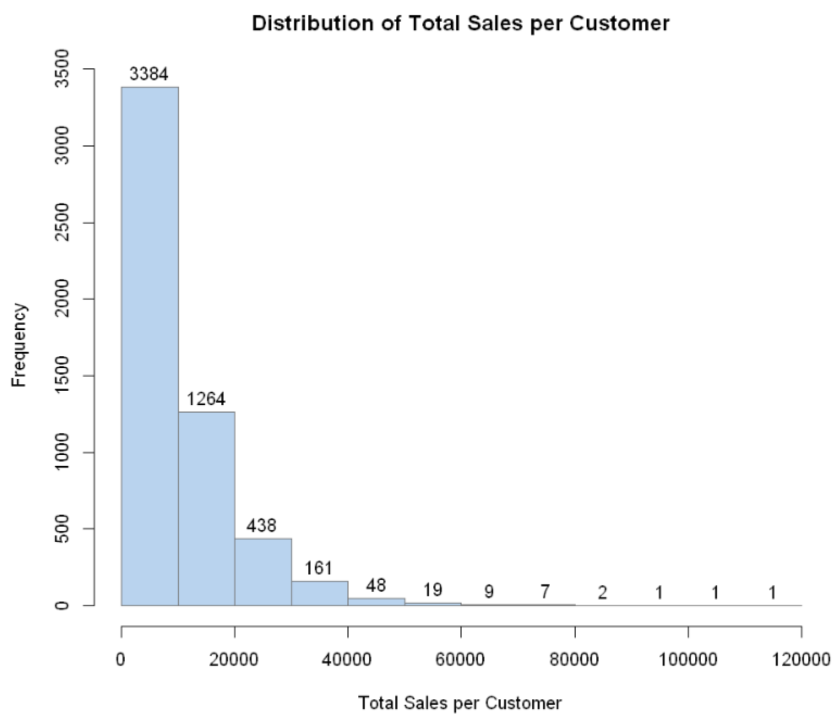


Figure 6: Distribution of Total Sales per Customer

3.3.3 Distribution of Inter-transaction Times

Here we can see the distribution of the inter-transaction times for all transactions in the dataset. I have limited the x-axis at 30 because of the presence of some outliers. However, these outliers are important as they are indicative of the difficult in accurately predicting a customer's future activity. In 545 cases we observe a waiting time of over 365 days. Some customers remain very inactive for long periods of time, and yet still make another transaction in the future. These types of customers will be very difficult to model since they don't defect but simply "hibernate" until the time comes when they transact again.

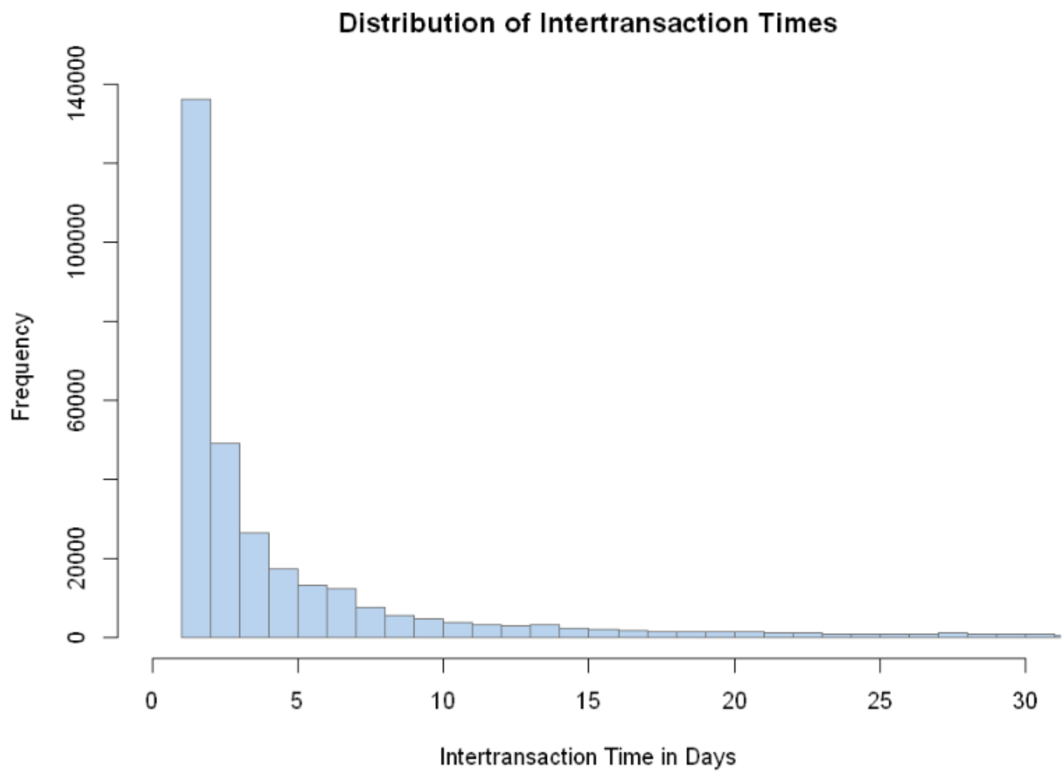


Figure 7: Distribution of Inter-transaction Times

4 Theory & Methods

4.1 Pareto NBD Model

4.1.1 Background

In 1987, Schmittlein, Morrison and Colombo first introduced the Pareto / NBD model (Schmittlein et al. 1987). This was developed to build upon research previously conducted by Andrew Ehrenberg, who had published his seminal article "The Pattern of Consumer Purchase" years earlier (Ehrenberg, 1959).

Ehrenberg used a negative binomial distribution (NBD) as a probability distribution for count data of sales of non-durable consumer goods. Customers purchase goods according to a Poisson process, he argued, and not all customers purchase at the same rate. Thus, he proposed that the heterogeneity in purchase rates across the customer base was distributed according to a Gamma distribution. Hence, his model is referred to as the NBD model, which is precisely the theoretical distribution of the combination of the Gamma-Poisson mixture.

However, the major drawback of the NBD model is its assumption that all customers remain active for the duration of the observation period. It does not account for the possible dropout of customers. Schmittlein et al. were attempting to address this issue, the non-observable dropout process. A customer may decide to bring their business elsewhere at any time, unbeknownst to the service provider. Given that the organization will typically not be notified of that customers' defection, they must model customer behaviour to understand the current activity of each customer. Considering the various timing patterns of customers (figure 3), we can appreciate the difficulty in assessing whether a customer is still alive or not, let alone building a stochastic parametric model.

The Pareto/NBD model was designed to allow for a customer dropout process while also modelling individual customer purchase frequencies in the same manner as Ehrenberg. It assumes an individual stochastic defection process for each customer and assumes that each customers lifetime is exponentially distributed with death rate μ.

4.1.2 Assumptions

The following is a summary of the assumptions of the Pareto/NBD regarding consumer behaviour, as outlined in Fader & Hardie (2005):

4.1.2.1 Assumption 1

Customers go through two stages in their "lifetime" with a specific firm: they are "alive" for some period of time, and then become permanently inactive.

4.1.2.2 Assumption 2

While alive, the number of transactions made by a customer follows a Poisson process with transaction rate λ; therefore, the probability of observing x transactions in the time interval $(0, t]$ is given by:

$$P(X(t) = x, \lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \forall x = 0,1,2 \ldots \tag{2}$$

This is equivalent to assuming that the time between transactions is distributed exponential with transaction rate λ,

$$f\left(t_j - t_{j-1}\middle|\, \lambda\right) = \lambda e^{-\lambda(t_j - t_{j-1})}, \forall t_j > \; t_{j-1} > 0 \qquad (3)$$

This postulates that the number of transactions follows a Poisson process which is equivalent to assuming that the waiting time between transactions follows an exponential distribution. The Poisson process with rate parameter λ is the count process for a timing process with independent exponentially distributed waiting times with mean $\frac{1}{\lambda}$. (Chatfeld and Goodhardt, 1973).

The exponential distribution is of course a special case of the gamma distribution, with the shape parameter being set to 1. Exponentially distributed variables are *memoryless*, which means that any information about the last event does not influence any event occurring in the immediate future.

$$P(T > s + t \,|T > s) = P(T > t) \; for \; al \; s, t \; \geq 0 \qquad (4)$$

This implies that the timing of a transaction does not depend on when the last purchase took place. Intuitively, one might think that customers purchase goods with some degree of regularity. For instance, a customer will replenish their stock of a certain item when they have used it and need more. The memoryless also implies that immediately after the initial purchase is the most likely time for the subsequent purchase (Morrison and Schmittlein, 1988). This memoryless property does not necessarily hold when considering purchasing consumer goods, as eluded to previously. This is an issue that we will consider later in the dissertation.

4.1.2.3 Assumption 3

Customer's unobserved "lifetime" of length τ (after which they are viewed as being inactive) is exponentially distributed with dropout rate μ:

$$f(\tau|\mu) = \mu e^{\mu\tau} \qquad (5)$$

This assigns an exponentially distributed lifetime with a "death" rate μ for each customer. Schmittlein et al. (1987) justify this since "the events that could trigger death (a move, a financial setback, a lifestyle change, etc.) may arrive in a Poisson manner". Logically, this seems like a reasonable assumption. However, the customers actual defection is not observable and hence is not possible to verify it. Furthermore, even if it was possible to verify the defection of a single customer, it occurs just a single time and reveals very little about the underlying death rate parameter μ.

4.1.2.4 Assumption 4

The prior distribution of the λ follows a gamma distribution with shape parameter r and scale parameter α:

$$g(\lambda|r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)} \qquad (6)$$

Here we assign a Gamma distribution to account for the heterogeneity in purchasing rates across customers and to allow for flexibility within the model, given the number of possible shapes of the two-parameter Gamma probability distribution. Aside from the extra flexibility and its positive skewness, Ehrenberg (1959) provides no other justification as to the choice of the Gamma distribution for this purpose. Ehrenberg explicitly models the heterogeneity across customers for modelling on the individual level. Therefore, he makes use of the information available about the entire customer base to form an estimate of the distribution of the parameters across customers. Schmittlein et al. (1987) state that "while there is not enough information to reliably estimate [the purchase rate] for each person, there will generally be enough to estimate the distribution of [it] over customers. [..] This approach, estimating a prior distribution from the available data, is usually called an empirical Bayes method".

4.1.2.5 Assumption 5

The prior distribution of the μ follows a gamma distribution with shape parameter s and scale parameter β:

$$g(\mu|s,\beta) = \frac{\beta^s \mu^{s-1} e^{-\mu\beta}}{\Gamma(s)} \tag{7}$$

This assumes that the rate parameter μ is Gamma distributed across the customer cohort. This Gamma distribution will have different parameters to the previous Gamma distribution. Similar to the modelling of the heterogeneity of the purchasing rate, this allows us to estimate the distribution of the death rate across the customer base and ultimately estimate the parameters for a single customer. This Gamma-Exponential mixture results in the Pareto distribution. As such, the overall model is termed the Pareto/NBD model.

4.1.2.6 Assumption 6

The transaction rate λ and the dropout rate μ vary independently across customers. For example, one customer who purchases very often is not assumed to have a longer or shorter lifetime expectancy than a customer who purchases less often. This is a necessary assumption to simplify the complex mathematical derivations of the model. Schmittlein et al. (1987) provide some reasoning for this assumption and Abe (2008) present some statistical evidence that λ and μ are indeed uncorrelated.

4.1.3 Model Specification

In general, we let Θ denote the parameters of a model. Then Bayes rule states that the posterior is proportional to the product of the prior and the likelihood:

$$\pi(\Theta|Data) \propto L(Data|\Theta)\,\pi(\Theta) \tag{8}$$

Where $L(Data|\Theta)$ denotes the likelihood, $\pi(\Theta)$ denotes the prior and $\pi(\Theta|Data)$ denotes the posterior.

The Pareto/NBD model is therefore:

$$\begin{aligned}\pi(\mathrm{r},\alpha,\mathrm{s},\beta,\{\lambda_i\},\{\mu_i\}|Data_i) \\ \propto \prod_i L(\,Data_i \mid \lambda_i,\mu_i) g(\lambda_i \mid r,\alpha) \mathrm{g}(\mu_i \mid s,\beta) \text{ x } \pi(\mathrm{r})\pi(\alpha)\pi(\mathrm{s})\pi(\beta)\end{aligned} \tag{9}$$

$\pi(\mathrm{r})$, $\pi(\alpha)$, $\pi(\mathrm{s})$ and $\pi(\beta)$ are the prior distributions on the parameters of the Gamma distribution of λ and μ. These are generally designed to be uninformative so that they have minimal influence on the posterior distribution. $Data_i$ refers to customer i's purchasing history.

4.1.4 Deriving the Likelihood Function Conditional on λ and μ

This derivation is outlined by Fader and Hardie (2005). When each of a customer's x transactions occurred during the period $(0, T]$ we denote these times by $t_1, t_2.., t_x$:

There are two possible ways this pattern of transactions could arise:

1. The customer is still alive at the end of the observation period $(i.e.\ \tau > T)$, in which case the individual-level likelihood function is simply the product of the (inter-transaction-time) exponential density functions and the associated survivor function:

$$\begin{aligned} L(\lambda \mid t_1, .., t_x, T, \tau > T) &= \lambda e^{-\lambda t_1} \lambda e^{-\lambda (t_2 - t_1)} \ldots \lambda e^{-\lambda (t_x - t_{x-1})} e^{-\lambda (T - t_x)} \\ &= \lambda^x e^{-\lambda T} \end{aligned}$$

2. The customer became inactive at some time τ in the interval $(t_x, T]$, in which case the individual-level likelihood function is:

$$\begin{aligned} L(\lambda \mid t_1, .., t_x, T, inactive\ at\ \tau \in (t_x, T]) &= \lambda e^{-\lambda t_1} \lambda e^{-\lambda (t_2 - t_1)} \ldots \lambda e^{-\lambda (t_x - t_{x-1})} e^{-\lambda (\tau - t_x)} \\ &= \lambda^x e^{-\lambda \tau} \end{aligned}$$

We note that in both cases, information on when each of the x transactions occurred is not required; In this case, t_x and x are sufficient summaries of a customer's transaction history. Using direct marketing terminology, t_x is recency and x is frequency. Removing the conditioning on τ yields the following expression for the individual-level likelihood function:

$$\begin{aligned} L(\lambda, \mu \mid x, t_x, T) &= L(\lambda | x, T, \tau > T) P(\tau > T | \mu) + \int_{t_x}^{T} L(\lambda \mid x, T, inactive\ at\ \tau \in (t_x, T]) f(\tau \mid \mu) d\tau \\ &= \lambda^x e^{-\lambda T} e^{-\mu T} + \lambda^x \int_{t_x}^{T} e^{-\lambda T} \mu e^{-\mu \tau} d\tau \\ &= \lambda^x e^{-(\lambda + \mu) T} + \frac{\lambda^x \mu}{\lambda + \mu} e^{-(\lambda + \mu) t_x} - \frac{\lambda^x \mu}{\lambda + \mu} e^{-(\lambda + \mu) T} \\ &= \frac{\lambda^x \mu}{\lambda + \mu} e^{-(\lambda + \mu) t_x} + \frac{\lambda^{x+1}}{\lambda + \mu} e^{-(\lambda + \mu) T} \\ &= \frac{\lambda^x}{\lambda + \mu} (\mu e^{-(\lambda + \mu) t_x} + \lambda e^{-(\lambda + \mu) T}) \end{aligned}$$

4.1.5 The Gibbs Sampler for the Pareto/NBD model

This likelihood distribution is intractable and hence very difficult to sample from, unless we use a technique such as Markov Chain Monte Carlo (MCMC). This is a simulation-based estimation procedure whereby random draws are recursively estimated from the full conditional distributions of the model and are used as conditioning arguments in subsequent draws. Upon convergence, these draws are samples from the true posterior distribution. We can verify that the MCMC chain has converged by inspecting trace plots which show the value that the parameter is taking for each iteration in the chain. We run multiple different chains, and if the values from each chain eventually converge and stay around the same level, we conclude that we are sampling from the stationary distribution of the parameter. In doing this, we generate:

1. Estimated marginal posterior distributions rather than point estimates
2. Individual-level parameter estimates
3. Straightforward simulations of customer-level metrics

We use the Gibbs sampler for MCMC inference of this Pareto/NBD model. As explained earlier, this is achieved by recursively generating draws from each of the densities in the following algorithm, which is outlined in Ma and Liu (2007):

1) Generate λ_i (one at a time for each customer):

$$\pi(\lambda_i \mid x_i, t_{xi}, T_i, \mu_i, r, \alpha) \;\propto\; \frac{{\lambda_i}^{x_i+r-1} e^{-\lambda_i \alpha}}{\lambda_i + \mu_i} (\mu_i e^{-(\lambda_i+\mu_i)t_{xi}} + \lambda_i e^{-(\lambda_i+\mu_i)T_i}) \qquad (10)$$

2) Generate μ_i (one at a time for each customer):

$$\pi(\mu_i \mid x_i, t_{xi}, T_i, \lambda_i, s, \beta) \;\propto\; \frac{{\lambda_i}^{x_i+s-1} e^{-\mu_i \beta}}{\lambda_i + \mu_i} (\mu_i e^{-(\lambda_i+\mu_i)t_{xi}} + \lambda_i e^{-(\lambda_i+\mu_i)T_i}) \qquad (11)$$

3) Generate α:

$$\pi(\alpha \mid \{\lambda_i\}, r) \;\propto\; \prod_{i=1}^{m} \alpha^r \, e^{-\lambda_i \alpha} \; x \; \pi(\alpha) \qquad (12)$$

Where m is the number of customers

4) Generate r:

$$\pi(r \mid \{\lambda_i\}, \alpha) \;\propto\; \prod_{i=1}^{m} \frac{\alpha^r {\lambda_i}^{r-1} e^{-\lambda_i \alpha}}{\Gamma(r)} \; x \; \pi(r) \qquad (13)$$

5) Generate β:

$$\pi(\beta \mid \{\mu_i\}, s) \;\propto\; \prod_{i=1}^{m} \beta^s \, e^{-\mu_i \beta} \; x \; \pi(\beta) \qquad (14)$$

6) Generate s:

$$\pi(s \mid \{\mu_i\}, \beta) \;\propto\; \prod_{i=1}^{m} \frac{\beta^s {\mu_i}^{s-1} e^{-\mu_i \beta}}{\Gamma(s)} \; x \; \pi(s) \qquad (15)$$

4.1.6 Analytical Workflow

The data requirements to implement the Pareto/NBD model are quite few which are outlined in Michael Platzer (2016). We require four pieces of information for each customer:

1. Frequency: How many transactions they made in the calibration period
2. Recency: The time of their last transaction
3. Monetary: The value of each transaction
4. The total time for which they were observed

The term RFM data is derived from points 1-3 above. The analysis starts by collecting a complete log of all transactions of the customer cohort. In our case, this refers to every purchase the individual customer made using their phone credit for the period in question (Feb 2017 to Feb 2019). The form of the data required is straightforward. Each row simply must contain a customer identifying field, the date and the amount of the transaction.

The next step is to convert the data into a customer-by-sufficient-statistic (CBS) summary table. This is simply a table with the data points necessary for the model to be run. Each row in the CBS refers to a specific customer and contains the required RFM data.

This analysis is carried out using the programming language r. Michael Platzer developed an r package called BTYDplus, which can apply, among others, the Pareto/NBD statistical model to describe and predict purchasing behaviour of customers in non-contractual settings. This package has been used to fit these models and I will reference it as I describe the analysis.

4.1.7 Fitting the Pareto/NBD Model

As mentioned above, this implementation of the Pareto/NBD model relies on MCMC simulation for parameter estimation. The first step is to draw the parameters of the model, namely r, α, s and β.

The MCMC process consists of building a Markov Chain that has the desired target distribution (posterior) as its stationary distribution. The algorithm performs random walks on this Markov chain and will eventually converge to the stationary distribution at which point the model will be sampling from the correct posterior distribution. We allow for an initial burn-in stage, during which we “throw away” the samples as the model has not yet converged and therefore the samples are not from the true posterior distribution. We also use what is known as a thinning interval, which means that we only consider every x-th sample in the chain. This is due to the high autocorrelation between consecutive iteration steps in the MCMC chain.

This MCMC chain was constructed to produce 100 samples from the stationary distribution of each parameter with 5000 MCMC steps after a burn in period of 500 steps and a thinning interval of 50. This was completed using the pnbd.mcmc.DrawParameters() function from the BTYDplus package.

Let’s inspect the trace plots of each parameter in figure 8 to verify that the MCMC chains have converged. I ran four chains and as we can see, each of the four chains converge around the same value for the parameter indicating that we have found the correct target posterior distribution.

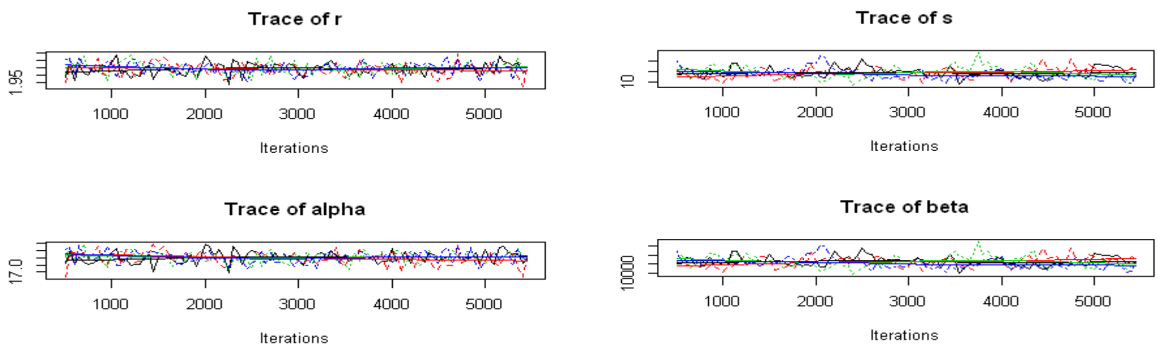


Figure 8: Trace Plots of Pareto/NBD Parameters

We then take the median of these samples to produce the following parameter estimates, with density plots below in figure 9:

	r	α	s	β
Median Estimate	2.04	18.01	14.30	16,163.36
2.5% Credible Interval	1.96	17.20	10.51	11,759.37
97.5% Credible Interval	2.12	18.77	19.38	21,934.07

Table 3: Pareto/NBD Parameter Estimates

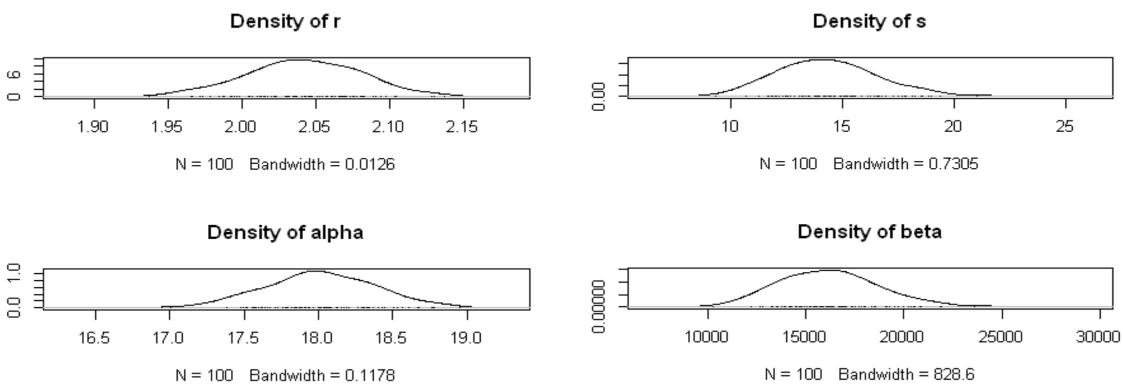


Figure 9: Density Plots of Pareto/NBD Parameters

4.1.8 Distribution of Transaction Rate

We can view the distribution of this gamma variable:

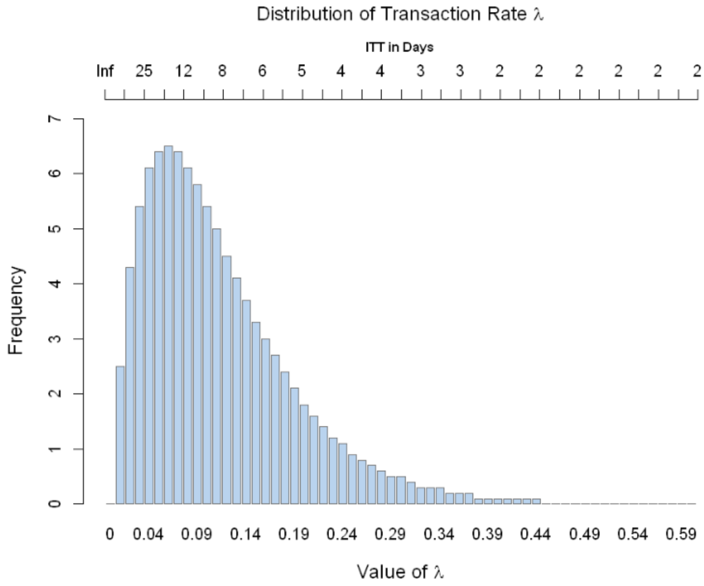


Figure 10: Distribution of Transaction Rate

Here we can see the distribution of the rate parameter λ on the lower x-axis, and the equivalent number of days inter-transaction time (ITT) on the secondary x-axis above the chart.

The mean estimate for the transaction rate λ is equal to $r\ /\ a\ =\ 0.113$. This equates to a mean ITT of $1\ /\ \lambda\ =\ 8.82\ days$ with a standard deviation of $r\ /\ (\sqrt{a})\ =\ 12.60\ days$.

This roughly agrees with our findings from the exploratory data analysis, which indicated an average inter-transaction time of ~10 days with a standard deviation of 11.44 days.

4.1.9 Distribution of Dropout Rate

We can view the distribution of this gamma variable:

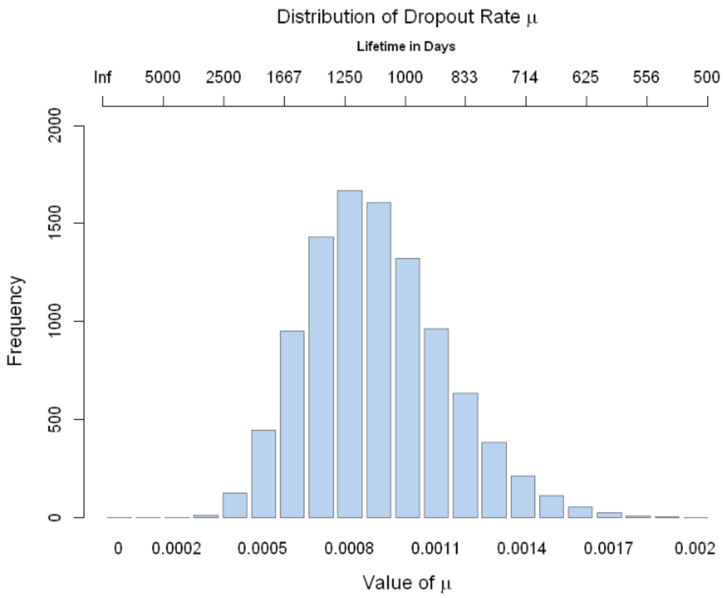


Figure 11: Distribution of Dropout Rate

Here we can see the distribution of the rate parameter μ on the initial x-axis, and the equivalent number of customer lifetime days on the secondary x-axis (blue).

The mean estimate for the dropout rate μ is equal to $\frac{1}{\beta/_{S}} = 0.000889$. This equates to a mean lifetime of $1/_{\mu} == 1{,}130$ days with a standard deviation of $\beta/_{\sqrt{S}} == 4{,}274$ days.

The expected lifetime of these customers is long. In general, mobile users tend to remain with the same mobile provider for a long period of time, and therefore our mean estimate of 3.1 years seems reasonable. The model is only trained on two years' worth of transaction history, and hence it is extrapolating into the future with these predictions. Given that we only have data on customers over a 24-month period, the maximum observed lifetime for a customer is 24 months. This is where the discrepancy between our average observed customer lifetime from the exploratory data analysis and this average predicted customer lifetime comes from.

4.1.10 Forecast Accuracy

To evaluate the forecast accuracy of our model we need to divide our dataset into a calibration period and a validation period. We estimate the parameters of the model using the calibration period and then we predict the number of transactions in the validation period. The differences between the predicted and actual values will inform our performance metrics. As mentioned earlier, we use the first 18 months of data, 1st-Feb-2017 to 31st-July-2018 as the calibration period and the final 6 months, 1st-Aug-2018 to 1st-Feb-2019 is considered as the validation period.

4.1.10.1 Cumulative Repeat Transactions

Regarding the initial questions we wanted to address, one of these was "How many transactions can we expect from our customers in the future?". Our modelling framework primarily predicts the number of transactions on an individual level, but we would expect the cumulative numbers to agree strongly with the overall transaction volume.

	Holdout
Actuals	63,248
NBD Model	93,945
Pareto/NBD Model	55,955

Table 4: Predicted No. Transactions Holdout Period

We can see that the Pareto/NBD model does a good job in predicting the total number of transactions during the holdout period. Of the 63,248 transactions that did occur, our model predicted 55,955. We will explore the components of the 12% difference in the later sections.

I also applied the original NBD to the dataset to act as a performance benchmark. The NBD model vastly overestimates the total number of transactions. Due to the lack of a dropout mechanism, the NBD model assumes all customers remain active throughout the validation period. Since no customers drop out, it assumes they keep purchasing which leads to this poor prediction.

4.1.10.2 Grouped by Transaction Count

Fader, Hardie & Lee (2005) proposed a performance evaluation method whereby we group customers based on their number of transactions during the 18-month calibration period. For each group, we compare the average predicted number of transactions during the validation period to the actual number of transactions in the holdout period.

In table 5 below, I have grouped the data by the count of transactions during the calibration period in groups of 10 transactions. We can then compare the actual number of transactions in the holdout period versus the expected number as predicted by the NBD model and the PNBD model. As the number of transactions in the calibration period increases, the relative accuracy of the PNBD model versus the NBD model begins to show. This can be clearly seen in figure 12 below. The model makes more accurate predictions when it has more data to learn from.

I have also included 95% Bayesian credible intervals which provides an interval within which our predicted number of transactions falls with probability equal to 95%.

Transactions in Calibration	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**	**100+**
Actuals	177	57	61	79	99	112	119	146	179	175	177
Expected NBD	22	58	92	128	165	199	234	270	306	343	22
Expected PNBD	8	19	37	60	82	111	129	149	191	215	8
Expected PNBD – 2.5%	0	0	0	1	8	8	11	11	19	19	0
Expected PNBD – 97.5%	34	61	90	127	158	200	231	267	324	345	34
Group Size	352	1117	825	689	488	354	310	233	201	159	352

Table 5: Predicted No. Transactions Holdout Period – Grouped by No. Transactions Calibration Period

We can visualize these numbers in a chart, as below.

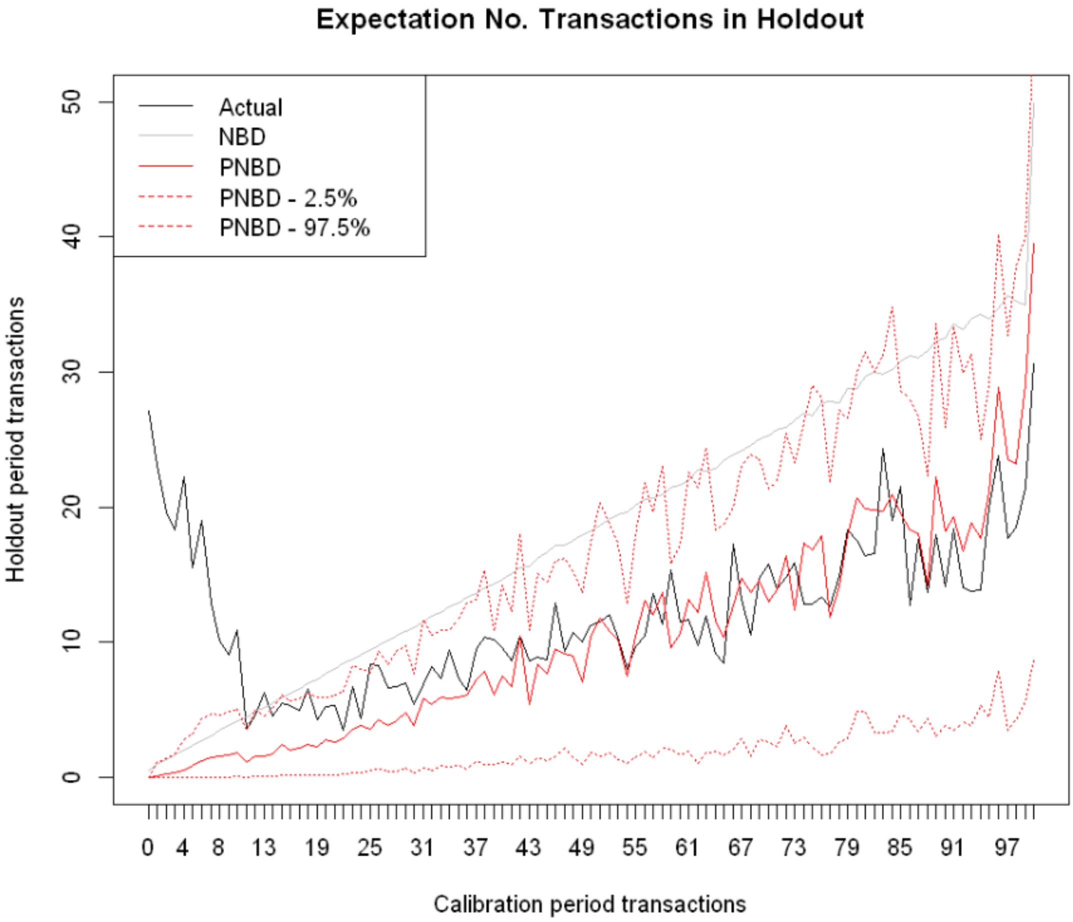


Figure 12: Expected No. Transactions in Holdout grouped by No. Transactions in Calibration

We can see that as the frequency in the calibration period increases, the PNBD model is much more accurate in its predictions for the holdout period. Intuitively, the model's predictive accuracy increases when there is more past data to analyze. In comparison, the NBD model predicts a linear relationship between calibration period transactions and holdout period transactions. As mentioned, the NBD model assumes each customer remains active and thus will assume that they transact as frequently during the holdout period as the calibration period.

4.1.10.3 Individual Level Forecasts

The primary objective of these models is to make conditional estimates for each individual customer based on their past transactional history. Naturally, each estimate will deviate from the correct figure. To evaluate our model accuracy, we must produce a single figure to represent the aggregation of these errors.

I have used three different metrics to evaluate the forecasts of each model. The first of these is known as Mean Absolute Error (MAE). MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.[1]

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{16}$$

The second metric considered is known as Mean Squared Logarithmic Error (MSLE). This can be interpreted as the measure of the ratio between the true and the predicted values. The logarithm makes the MSLE only care about the relative differences between the true and predicted values. This means that MSLE will treat small differences between small true and predicted values approximately the same as big differences between large true and predicted values.[2]

$$MSLE = \frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \tag{17}$$

The final metric considered is known as Root Mean Squared Error (RMSE). RMSE is the standard deviation of the residuals. Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are, in other words, it tells you how concentrated the data is around the line of best fit. [3]

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{28}$$

[1] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d
[2] https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-squared-logarithmic-error
[3] https://www.statisticshowto.datasciencecentral.com/rmse/

	MAE	MSLE	RMSE
NBD Model	1.09	2.83	17.22
Pareto/NBD Model	0.76	1.88	14.88

Table 6: Error Metrics for Pareto/NBD model

As expected, on an individual customer level the Pareto/NBD model greatly outperforms the NBD model with respect to each of the evaluation metrics.

4.1.11 Analysis of Model Errors

Here, we attempt to understand the where our modelling approach could be improved. We will view the transaction timing patterns for the calibration versus validation period for two subsets of customers. The first subset will be the 10 customers for which our model worst underestimated the number of holdout transactions. The second subset will be the 10 customers for which our model worst overestimated the number of holdout transactions.

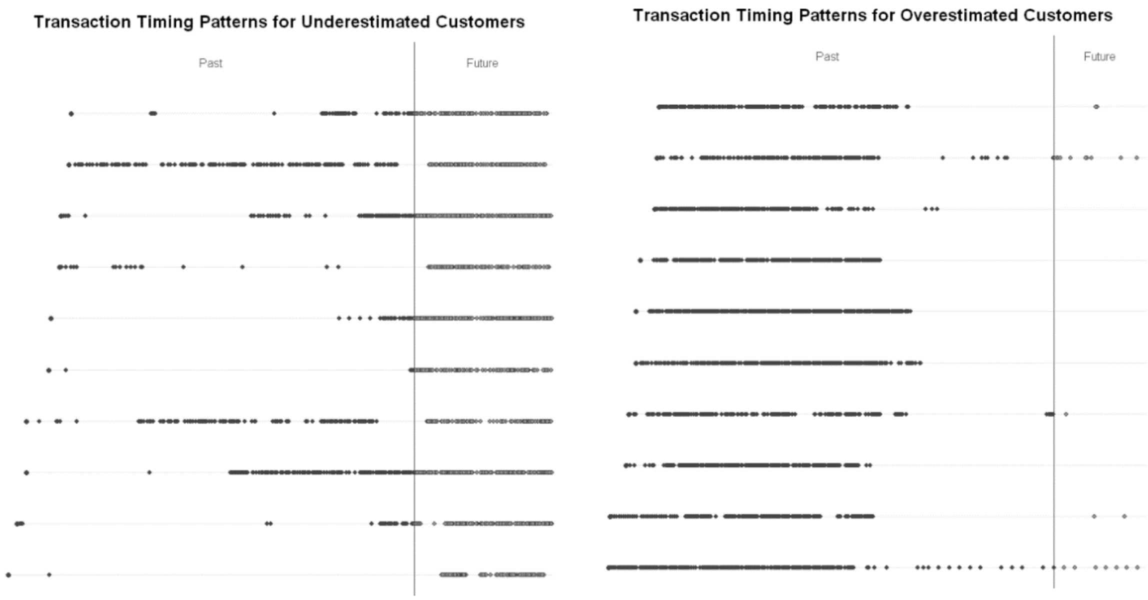


Figure 13: Transaction Timing Patterns – Worst Underestimated & Worst Overestimated Customers

Most of the underestimated customers had quite few transactions during the calibration period. As such, it would be very difficult for our model to understand and predict that they would proceed to make so many transactions in the holdout period. This type of behaviour is very unpredictable and as such there is nothing that we could do to improve our model in this regard.

With respect to the second figure, these customers seem to have a lot of regularity in their purchasing behaviour. Most of these customers seemed to stop purchasing mobile plans during the calibration period. Our stochastic model has not picked up on this since we overestimated the number of purchases in the holdout period. These customers seem to purchase mobile plans in specific intervals, and the sudden stop in purchasing during the calibration phase would indicate that they have churned. Perhaps they purchased new plans every week for several months and then changed network or moved to a new country. This is a sign of regularity in purchasing behaviour, which is something that we can include explicitly in our modelling framework.

4.2 Regularity

Through our analysis we have identified certain issues with our models specifically related to regularity in customer purchasing.

1. Intuitively, one would assume that customers purchase mobile plans in certain regular intervals. Mobile plans last for a certain amount of time before they need to be replenished.
2. For some customers, the time between succeeding transactions cannot be considered as random. It appears the transaction timing process follows a certain pattern for some customers. When we observed the plot of inter-transaction times in figure 13, we noticed that some customers seem to purchase in regular intervals.

Both models discussed so far maintain the assumption of a Poisson purchase process (i.e. NBD). As previously noted, the dropout process is latent, but the purchase process is directly observable and hence we can extract precise customer-level information. As such, a logical step is to focus on the purchase process and build a more flexible distribution to account for a larger range of real-world timing patterns.

Accounting for deviations from the Poisson purchasing in repeat buying models has been considered before in the marketing literature. For example, Herniter (1971) was the first to propose modelling inter-transaction times using an Erlang-k family of distributions. These allow for different degrees of regularity in timing patterns (i.e. higher k leads to stronger regularity). Chatfield and Goodhardt (1973) combined the Erlang-k with a gamma distribution to reflect the variation in the purchase rate across customers and termed the resulting model a condensed negative binomial distribution (CNBD). However, Chatfield and Goodhardt worried that “the CNBD formulas are so much more complex that it is doubtful if the small improvements in fit that seem possible justify the extra effort.”

We can incorporate a Gamma distribution instead of the Erlang-k distribution to add extra flexibility in the modelling of the inter-transaction times. The Erlang-k distribution is effectively a special case of the gamma distribution with the shape parameter no longer restricted to integer values. This model is known as the Pareto/GGG model and it was developed by Platzer & Reutterer (2016). We will discuss this model in further detail in the next section.

4.2.1 Estimating Regularity

As mentioned above, customer purchasing patterns can either be completely random, i.e. following a Poisson process, or they can be deterministic. We need a measure of regularity to distinguish between these.

One method to determine this regularity factor, known as the maximum likelihood method as outlined in Platzer (2016), involves fitting separate Gamma distributions to the observed inter-transaction times of customers as long as they have at least 10 recorded transactions. Then we inspect the distribution of the shape parameters of the Gamma distributions. If the median shape parameter is close to 1, we can accept the Poisson assumption for the group of customers. If the median value is larger than 1, this could be a strong indicator for the presence of regularity for these customers.

Customer ID	Regularity
1	1.2789
2	1.5410
3	1.7551

Table 7: Regularity Estimate for 3 Customers

We apply this MLE method for estimating regularity to each customer in the cohort. As such, we place a figure on the regularity of each customer in the cohort.

In figure 14 we can see the distribution of regularity across the customer base:

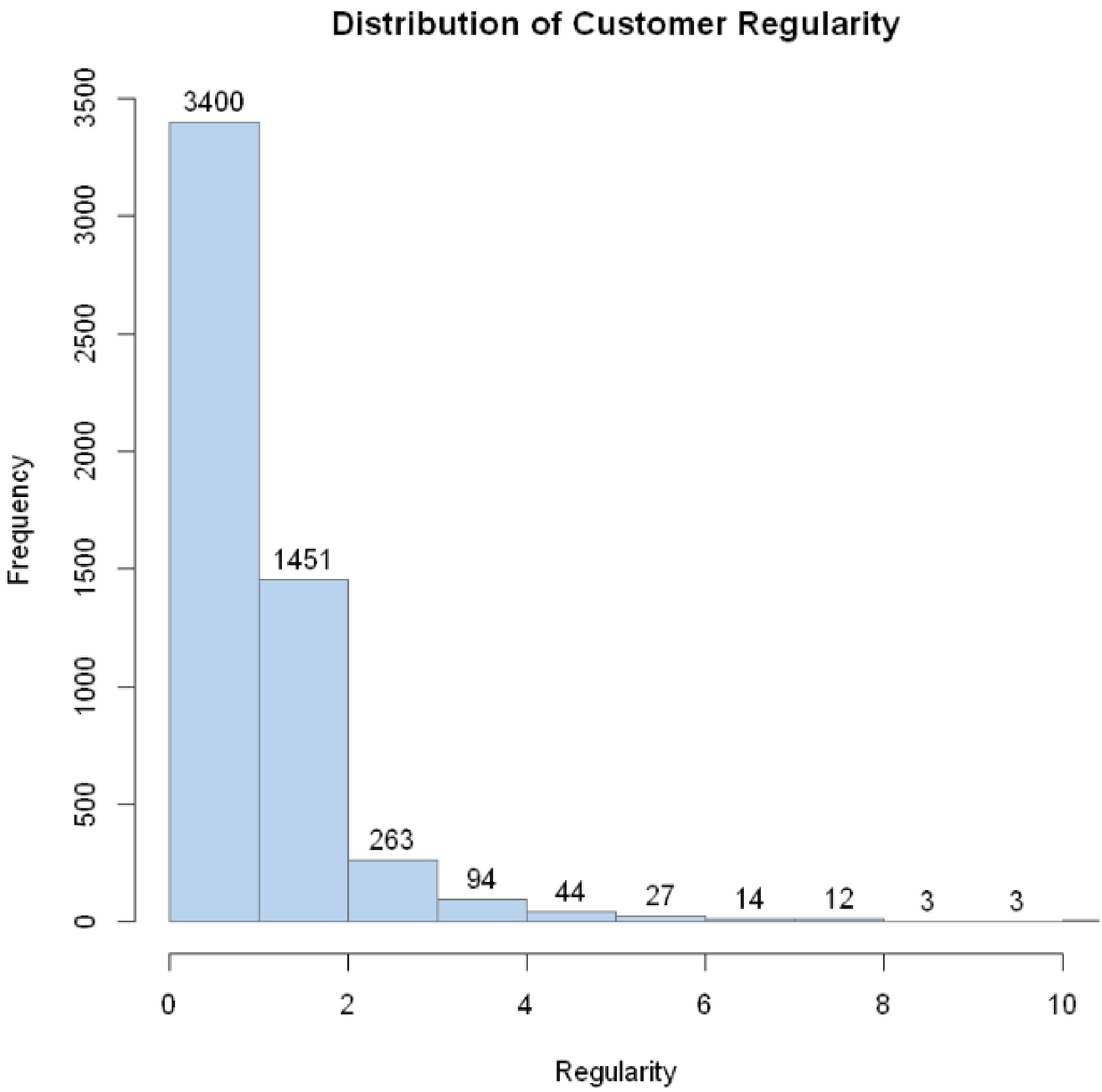


Figure 14: Distribution of Customer Regularity

As we can see, the distribution is heavily right-skewed, and the overwhelming majority of customers in this cohort do not purchase with strong regularity. However, we are only considering a relatively small cohort of customers and there will be many more customers showing signs of regularity in the real world. We will explore the possibility of applying a different model to the customers that do exhibit regularity in their purchasing patterns.

To do this, we must split the original dataset into two, one for customers exhibiting low levels of regularity and the other for customers exhibiting strong regularity. In this case, we would apply the Pareto/NBD model to the low regularity dataset, given its NBD assumption, and we would apply the Pareto/GGG model to the high regularity dataset.

4.2.1.1 Splitting the Dataset

In order to distinguish between customers that purchase with high regularity and those with low regularity, we need to decide upon a regularity threshold. To do this, I ran multiple test scenarios with different threshold values. As previously mentioned, a regularity value of 1 indicates no presence of regularity. As we saw in figure 14, there are few customers with high levels of regularity and as such I have set the upper limit as 2.6.

Hence, for threshold values between 1 and 2.6:

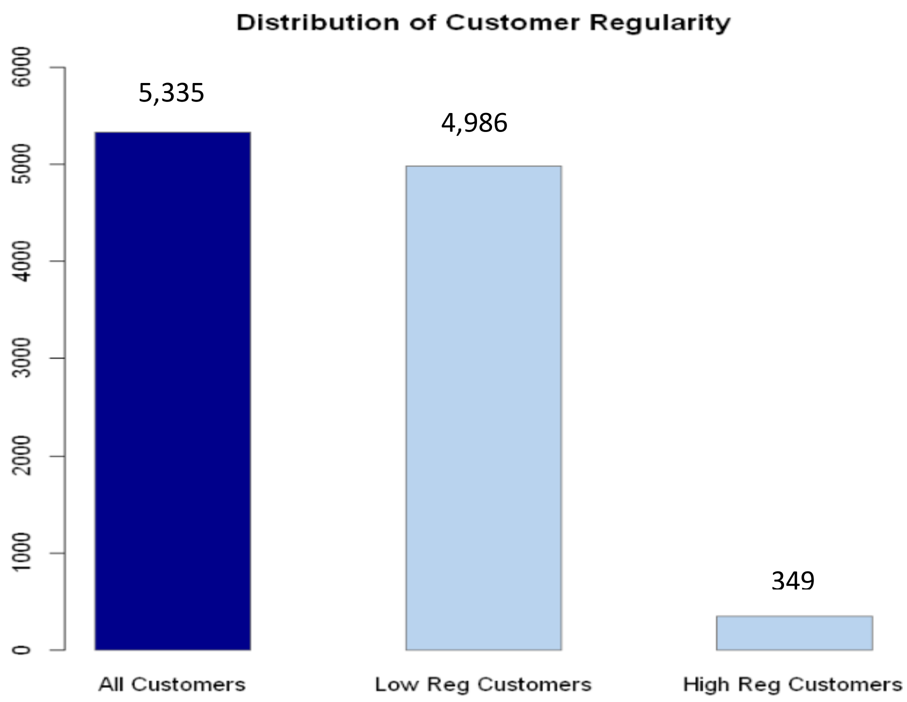
1. I took a random sample of 2500 customers from the low regularity dataset and a random sample of 290 customers from the high regularity dataset.
2. Fit the Pareto/NBD model to the low regularity dataset and predicted the number of transactions in the holdout period for each customer.
3. Fit the Pareto/GGG model to the high regularity dataset and predicted the number of transactions in the holdout period for each customer.
4. Summed the mean absolute error of predictions for each model.

Threshold	Total Error
1.0	1.5756
1.2	1.4611
1.4	1.3836
1.6	1.3588
1.8	1.2856
2.0	1.3033
2.2	1.2621
2.4	1.2217
2.6	1.2306

Table 8: Total Error for Threshold Values

The threshold value corresponding to the minimum overall error was selected as the optimum threshold value (2.4).

Applying this threshold to the dataset, we have the following summary statistics:



Here we can see a boxplot for the distribution of regularity among customers in the low regularity dataset and customers in the high regularity dataset.

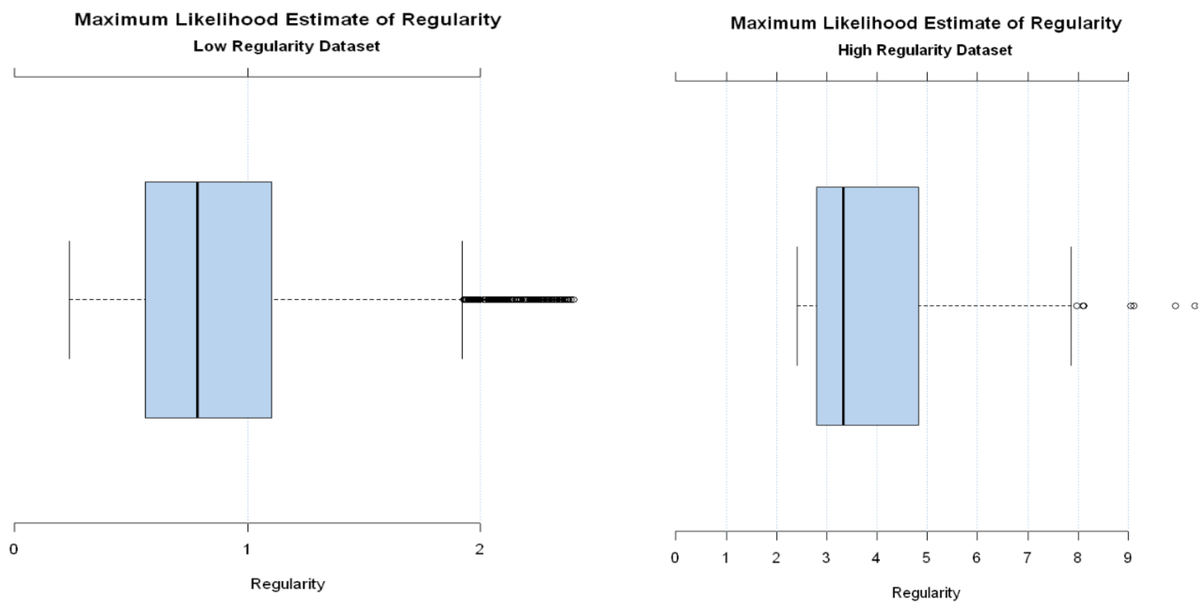


Figure 16: Boxplot of Regularity for Low Regularity Customer set vs High Regularity Customer set

As we can see, the customers in the low regularity dataset exhibit a median value of 0.78, well below 1 and thus exhibiting non-regular Poisson behaviour. On the other hand, customers in the high regularity dataset exhibit a median value of 3.33. This is a strong indicator of the presence of regularity in their purchasing behaviour.

4.3 Pareto/GGG Model

4.3.1 Background

This model was developed by Michael Platzer & Thomas Reutterer and is outlined in "Ticking Away the Moments: Timing Regularity Helps to Better Predict Customer Activity" (2016). It is an extension of the Pareto/NBD model as described above.

The major update was replacing the exponential distribution with a Gamma distribution to model customer's inter-transaction times. The shape parameter k of this Gamma distribution can therefore vary across the customer base according to a $gamma(t, \gamma)$ distribution. As a result, we now allow for heterogeneity in frequency, dropout and purchasing regularity. When there is a degree of regularity in the purchasing behaviour of the customer cohort, this Pareto/GGG can yield improved inferences about customers' future activity.

Using a hierarchical Bayesian framework, the models update a prior belief about a customer's regularity by considering the available information about the variance in inter-transaction times across the customer base. The model draws more accurate inferences about the customers regularity with more available transactions.

4.3.2 Assumptions

The following is a summary of the assumptions of the Pareto/GGG regarding consumer behaviour, as outlined in Platzer & Reutterer (2016):

4.3.2.1 Assumption 1

While the customer remains alive, their ITTs $\Delta t_j := t_j - t_{j-1}$ follow a gamma distribution, with shape parameter k and rate parameter $k\lambda$:

$$\Delta t_j \sim Gamma(k, k\lambda) \tag{19}$$

Here we are assuming that the customer's ITTs follow a gamma distribution. Recall that for the Pareto/NBD we modelled the number of transactions made by the customer as a Poisson process with transaction rate λ. This was equivalent to modelling the customers ITTs as exponential with transaction rate λ. As previously noted, this implies that the time immediately after a purchase is the most likely time for a repurchase. As noted above, this implication is problematic since these customers do show some signs of regularity in their purchasing habits. The Pareto/GGG model replaces this exponential distribution with a gamma distribution. The shape parameter can vary across customers, as noted in assumption 3.

4.3.2.2 Assumption 2

A customer remains alive for an exponentially distributed lifetime τ:

$$\tau \sim Exponential(\mu) \tag{20}$$

4.3.2.3 Assumption 3

The individual-level parameters $\{k, \lambda, \mu\}$ follow gamma distributions across customers independently:

$$k \sim Gamma(t, \gamma); \tag{21}$$

$$\lambda \sim Gamma(r, \alpha); \tag{22}$$

$$\mu \sim Gamma(s, \beta); \tag{23}$$

We are now allowing for heterogeneity in purchase frequency, customer dropout rate and regularity. Note that k is the newly introduced parameter which determines the regularity of the transaction timings.

4.3.3 Model Specification

To estimate the parameters of the Pareto/GGG, we begin by generating a full hierarchical Bayesian model with hyperpriors for the heterogeneity parameters, and then generate draws of the marginal posterior distributions using an MCMC sampling scheme.

This model has the same data requirements as the Pareto/NBD model, with the addition of a summary statistic of historic transaction timings, the sum over the logarithmic ITTs. This addition is easy to implement, as it can be derived from the same data inputs as required for the Pareto/NBD.

4.3.4 The Likelihood Function for the Pareto/GGG

As outlined in Platzer & Reutterer (2016), the likelihood of observing x inter-transaction times Δt_j, and then having no further transaction occur until time T, or in case of the customer churning, until time τ $(i.e.\ \Delta t_{x+1} > min(T, \tau) - t_x)$ can be expressed as follows:

$$\begin{aligned} L(k, \lambda \mid t_1, .., t_x, T, \tau) &= \left(\prod_{j=1}^{x} f_\Gamma\left(\Delta t_j \middle| k, k\lambda\right) \right) \left(1 - F_\Gamma(\min(T, \tau) - t_x | k, k\lambda)\right) \\ &= \frac{(k\lambda)^{kx}}{\Gamma(\mathrm{k})^x} e^{-k\lambda t_x} \left(\prod_{j=1}^{x} \left(\Delta t_j\right)^{k-1} \right) (1 - F_\Gamma(\min(T, \tau) - t_x | k, k\lambda)) \end{aligned}$$

4.3.5 Fitting the Pareto/GGG Model

To estimate the parameters for the Pareto/GGG model, we apply the function pggg.mcmc.DrawParameters() from the BTYDplus package in r. In addition to the RFM data requirements of the Pareto/NBD model, the Pareto/GGG model requires one additional requirement, the sum over the logarithmic ITTs. This is required to estimate regularity of each customer. Since we already have the complete log of transactions for each customer, we do not need to collect any extra data for its implementation.

We rely on MCMC to construct a markov chain and to set the desired posterior as the stationary distribution of each parameter. The MCMC chain was constructed to produce 100 samples from the stationary distribution of each parameter with 5000 MCMC steps after a burn in period of 500 steps and a thinning interval of 50. We will inspect the trace plots to ensure the MCMC chains have converged for each individual parameter. As we can see, each of the four chains correlate very strongly and results in stable estimates for each parameter.

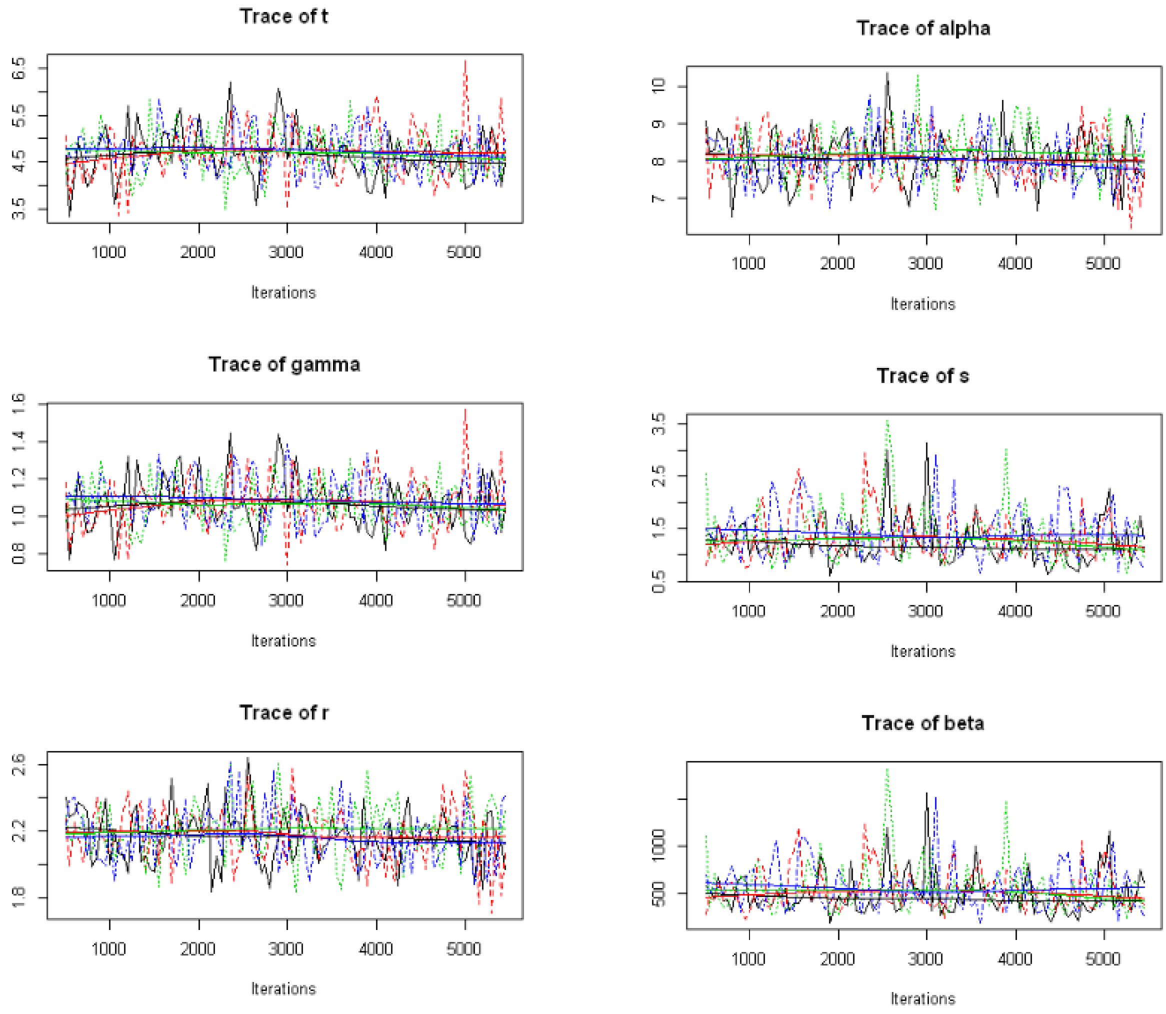


Figure 17: Trace Plots of Pareto/GGG Parameters

Now that we have our parameter samples, we pass these forward to draw the expected number of future transactions per customer. Based on each combination of parameter samples, the model predicts a different number of future transactions. We take the median of these predictions as our final predicted number of transactions for the given customer in the holdout period.

We want to compare the performance of the Pareto/GGG versus the performance of the Pareto/NBD model specifically on the “high regularity” dataset, to test whether it is more effective in modelling the expected number of transactions. As such, we do not need to fit the model to the entire dataset to understand the distribution of each parameter, as we did with the Pareto/NBD model. Instead, we proceed to dividing the dataset into a calibration and validation period and evaluating the forecast accuracy of each model on the high regularity dataset.

4.3.6 Forecast Accuracy

Using the same calibration and validation periods as for the Pareto/NBD model, we fit the Pareto/GGG model and estimate the parameters. Using these parameters, we can generate predictions for the total number of transactions in the holdout period.

4.3.6.1 Cumulative Repeat Transactions

Again, we are answering the question "How many transactions can we expect from our customers in the future?". We are comparing the predictive performance of the benchmark NBD model, the Pareto/NBD model and the new Pareto/GGG model.

	Holdout
Actuals – High Reg	5,591
NBD – High Reg	10,220
Pareto/NBD – High Reg	6,774
Pareto/GGG – High Reg	6,662

Table 9: Predicted Transactions in Holdout

Both the Pareto/NBD and the Pareto/GGG perform well here, but the Pareto/GGG is more accurate in predicting the aggregated cumulative repeat transactions predictions.

Note that since we are dealing with a much smaller set of customers, 349 versus the previous total of 5335, we do not have enough data to group by the number of calibration period transactions.

4.3.6.2 Individual Level Forecasts

Again, we will calculate the error performance metrics to compare the actual number of transactions in the holdout period versus the predicted number of transactions.

	MAE	**MLSE**	**RMSLE**
NBD – High Reg	0.9944	4.4654	22.2321
Pareto/NBD – High Reg	0.4046	0.3438	14.6037
Pareto/GGG – High Reg	0.3904	0.3353	14.3252

Table 10: Error Metrics for Models on High Regularity Dataset

The Pareto/GGG model outperforms the Pareto/NBD model here on a disaggregated level (i.e. on a customer level). It can leverage the regularity in customer purchasing behaviour and can predict the number of transactions that each customer will have in the holdout period more accurately than the Pareto/NBD model.

4.4 Combination Model: Pareto/NBD & Pareto/GGG

For the final model, we will split our dataset into two cohorts based on regularity in purchasing behaviour. We proceed to apply the Pareto/NBD model to the low regularity dataset and the Pareto/GGG model to the high regularity dataset. The combined results from this will be compared to separately applying the Pareto/NBD model and the Pareto/GGG model to the entire dataset. Therefore, we will be able to understand whether the combination of both models truly outperforms the models on an individual basis.

This dataset split will be based on the optimal regularity threshold as determined earlier in the analysis. We apply the same MCMC procedure to draw the parameters for both models and draw the expected number of future transactions based on these parameters. This is carried out for each model on the respective datasets. I then concatenated the CBS matrices together and created a new column representing the ultimate predicted number of transactions for the holdout period. From here, I implemented the same performance evaluation techniques.

Once I had estimated the model parameters, the models can be sued to make predictions regarding other types of customer behaviour. In particular, they can be used to calculate the probability that the customer will still be alive / active at the end of the forecasted period.

4.4.1 Forecast Accuracy

We used the very same calibration and validation periods as for each of our models.

4.4.1.1 Cumulative Repeat Transactions

Here we are comparing the results of the NBD model, the Pareto/NBD model and the combination of the Pareto/NBD & the Pareto/GGG models.

	Holdout
Actuals	63,248
NBD	93,945
Pareto/NBD	56,107
Pareto/GGG	55,986
Pareto/NBD & Pareto/GGG	56,369

Table 11: Predicted Number of Transactions in Holdout

As we can see, the combination of the Pareto/NBD model and the Pareto/GGG model produces the closest estimate to the true value.

4.4.1.2 Grouped by Transaction Count

We now group the customers based on their number of transactions during the 18-month calibration period. For each group, we compare the average predicted number of transactions during the validation period to the actual number of transactions in figure 18.

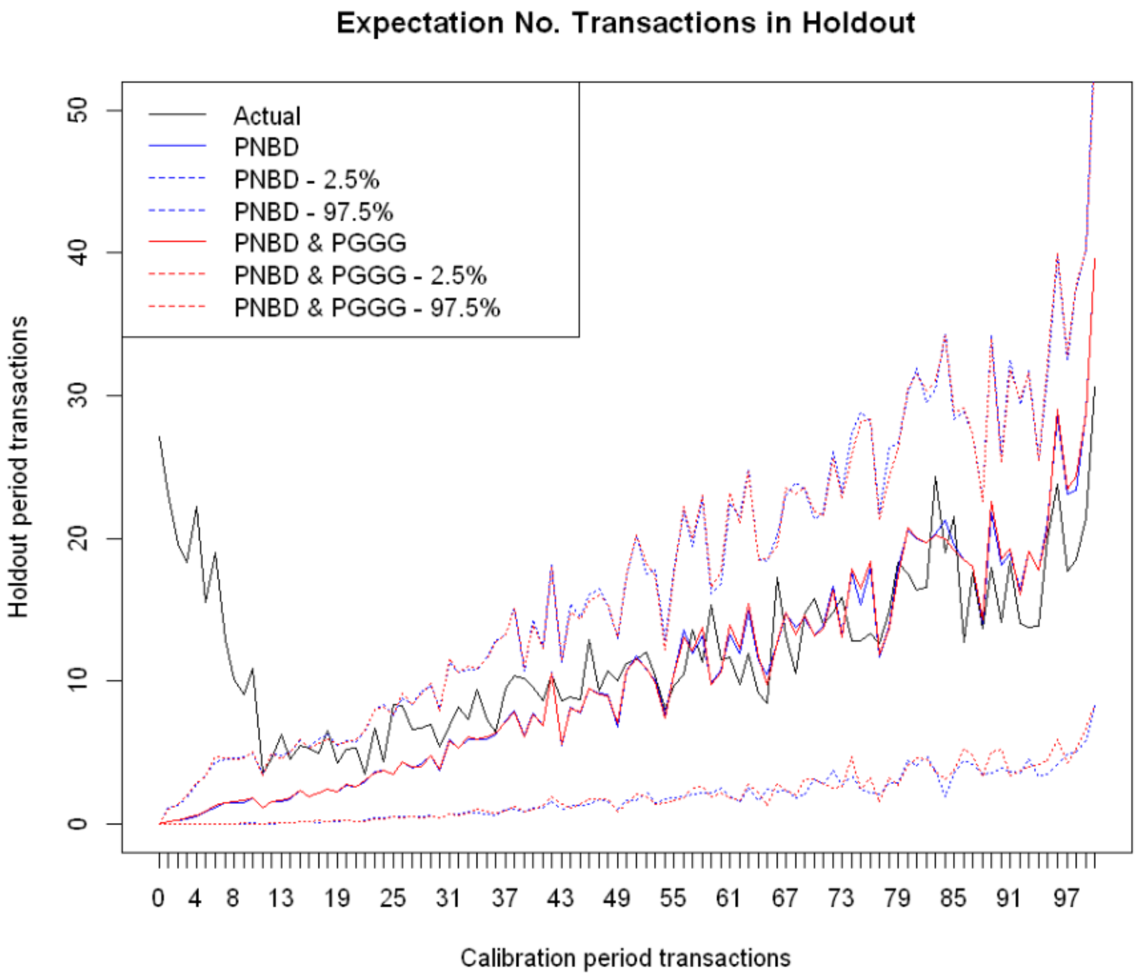


Figure 18: Expected No. Transactions in Holdout grouped by No. Transactions in Calibration

The results for the combination of both models versus the Pareto/NBD are very similar as can be seen above. Referring to figure 15, the majority of customers in this cohort do not exhibit regularity in their purchasing patterns and hence their output would not be changed in this model. Therefore, the updated modelling framework only affects a minority of customers and we would expect the lines to correlate very closely.

4.4.1.3 Individual Level Forecasts

We use the same performance metrics to compare the actual number of transactions in the holdout period versus the predicted number of transactions.

	MAE	MSLE	RMSE
NBD	1.086	2.833	17.219
Pareto/NBD	0.763	1.874	14.895
Pareto/GGG	0.762	1.885	14.883
Pareto/NBD & Pareto/GGG	0.761	1.864	14.853

Table 12: Error Metrics for each Model on Entire Dataset

The combination of both models performs slightly better as expected. This is only a slight improvement as the dataset is heavily weighted towards low regularity customers and therefore the “updated” model only affects a minority of customers.

4.5 Analysis of Combination Model versus Individual Models

As discussed previously, the Pareto/NBD model assumes a Poisson purchasing process which implies random purchasing times. In contrast, the Pareto/GGG model allows for regularity in consumer purchasing behaviour to improve inferences about customers’ latent activity status, Platzer & Reutterer (2016). In section 4.4.1 we compared the performance of three models on the entire dataset; Pareto/NBD, Pareto/GGG and the combination of the Pareto/NBD and Pareto/GGG on the low regularity dataset and the high regularity dataset respectively. Our results showed that the combination of both models outperformed the individual models.

The Pareto/NBD model is more effective at predicting future transactions than the Pareto/GGG when the customers exhibit low levels of regularity. This is because the Pareto/GGG replaces the NBD with a Gamma distribution. Once we confirm that the customers do not purchase with any degree of regularity, the Poisson assumption is proven to be true. As such, attempting to model these customers purchasing behaviour with a Gamma distribution would not be as effective as the NBD approach.

By dividing the dataset into two distinct groups, we can leverage the relative strengths of each model and produce more accurate forecasts as seen in the results in section 4.4.1.

4.6 Implementing the Final Model

Now that we have estimated the parameters for both models, we can use these to predict how many transactions each customer will make over the next defined period. We are going to consider the following 365 days for this analysis.

We sample the number of transactions during this 365-day period from the posterior distribution and take the median value of the samples as our prediction. We also make a prediction for the probability that the customer will still be alive after 365 days, and the probability that they will still be active. In this case, the probability of being active refers to the probability of making at least one transaction in the 365 period, and the probability of being alive is the probability of making another transaction at any time in the future, as outlined in Platzer (2016).

We also use this information to predict the overall value that the customer will bring to the company over this 365-day period. We simply take the product of the customers median sales amount per transaction during the observation period and their predicted number of transactions. This gives us the expected total revenue for the customer over the 365-day period.

4.6.1 Answering Managerial Level Questions using Final Model

Now that we have built the final model and made our predictions, we can address the managerial questions posed in the beginning of this thesis.

4.6.1.1 How many customers does the firm still have?

To answer this question, we would set a threshold for the probability of being alive and consider all customers above this threshold to be still alive at the end of the period.

	No. Customers Alive
Threshold = 0.95	2,075

Threshold = 0.80	2,600
Threshold = 0.60	2,799
Threshold = 0.40	2,938
Threshold = 0.20	3,119

Table 13: Number of Customers Alive

4.6.1.2 How many customers will be active in one year from now?

Similarly, we would set a threshold for the probability of being active and consider all customers above this threshold to be still active at the end of the period. We can see here that more customers tend to remain alive than active. This makes sense given the above definition of alive versus active.

	No. Customers Active
Threshold = 0.95	0
Threshold = 0.80	867
Threshold = 0.60	2,595
Threshold = 0.40	2,836
Threshold = 0.20	3,024

Table 14: Number of Customers Active

4.6.1.3 How many transactions can be expected in next 365 days?

We sum the number of expected transactions across all customers for the 365-day period.

Total Expected No. Transactions	90,287

Table 15: Expected Number of Transactions

4.6.1.4 Which customers can be considered to have churned?

We would set a threshold for the probability of being alive and consider all customers below this threshold to have churned during the period.

	No. Customers
Threshold = 0.2	2,212

Table 16: Number of Churned Customers

4.6.1.5 Which customers will provide the most value to the company going forward?

We rank each customer based on their predicted customer value over the 365-day period. Then we simply identify the top x number of customers in this ranked list.

5 Discussion and Conclusions

This dissertation was completed in coordination with the data analytics company to develop a framework to model the customer lifetime value of one of their largest clients operating in the Telecoms industry. After researching the literature regarding this subject, we discovered the Pareto/NBD family of models. The first of these that was applied was the Pareto/NBD model. This has been implemented many times before in various academic papers and commercial organizations. The major problematic implication of these models is the assumption of the Poisson process with respect to the transaction timings (i.e. random transaction timings). Nevertheless, this model has consistently performed well throughout the literature and has been successful in predicting count patterns in real world data. The Pareto/NBD model performs well on our overall dataset but as we discovered, can be improved upon.

On a disaggregated level, i.e. on a customer level, the importance of accurately predicting the timing of customers transactions is much higher. If we can understand the true current status of each customer, our predictions for their future transactions will be a lot more accurate. We analyzed the customers for which we overestimated their number of future transactions and discovered a pattern. These customers did not exhibit random purchasing times but seemed to purchase in regular intervals. These customers appeared to drop out during the calibration period, but our stochastic model did not pick up on this.

Therefore, we applied another model which was developed to account for regularity in customer purchasing. This is known as the Pareto/GGG model. This model has also been applied a lot throughout the literature. It is possible to estimate the degree of regularity in purchasing behaviour for each customer and subsequently divide the dataset into a high regularity dataset and a low regularity dataset. To assess the validity of the Pareto/GGG model, I tested the performance of the Pareto/NBD versus the Pareto/GGG on the high regularity dataset. The Pareto/GGG clearly outperformed the Pareto/NBD model. As such, I hypothesized that categorizing customers based on their regularity before deciding which modelling framework to apply could produce more accurate results.

In the available literature, the authors a single model to the entire dataset. In this dissertation, I propose a new approach. Choosing a suitable regularity threshold, we can divide the dataset into cohorts and estimate the parameters of the Pareto/NBD model on the low regularity dataset and estimate the parameters of the Pareto/GGG model on the high regularity dataset. Given that predictions are made on a customer level, we can then make predictions for each customer by applying the most suitable model.

The results of the Pareto/NBD model and the Pareto/GGG model versus the combination model prove this hypothesis to be correct. The combination of both models leverages the relative strengths of each model to achieve better predictions on an aggregated level as well as improved error metrics on a disaggregated level.

Once we had chosen the most effective modelling framework, it was straightforward to answer the managerial questions posed at the beginning of this dissertation. These predictions will provide valuable insights to the company regarding future customer behaviour.

6 References

Kotler and Armstrong, Principles of Marketing (1996), 7th ed., Englewood Cliffs, NJ: Prentice-Hall

Giuliano Rocco Tirenni, Allocation of Marketing Resources to Optimize Customer Equity (2005), University of St. Gallen

Floral & Friberg, Modelling Customer Lifetime Value in the Telecom Industry (2003), Lund University

D.C. Schmittlein, D.G. Morrison, and R. Colombo. Counting your customers: who are they and what will they do next? (1987), Management Science

A.S.C. Ehrenberg, The Pattern of Consumer Purchases (1959), Applied Statistics

Peter S. Fader & Bruce G. S. Hardie, A Note on Deriving the Pareto/NBD Model and Related Expressions (2005), URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.3165&rep=rep1&type=pdf

Morrison and Schmittlein, Generalizing the NBD Model for Customer Purchases: What are the Implications and Is it Worth the Effort? (1988), Journal of Business and Economic Statistics

Chatfeld and G.J. Goodhardt, A Consumer Purchasing Model with Er-lang Inter-Purchase Time, (1973), Journal of the American Statistical Association

Fader, Hardie & Lee, “Counting Your Customers” the Easy Way: An Alternative to the Pareto/NBD Model (2005), Marketing Science

Ma and Liu, The MCMC Approach for Solving the Pareto/NBD Model and Possible Extensions (2007), IEEE

Abe, "Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model (2008), Marketing Science

Michael Platzer, Customer Base Analysis with BTYDplus (2016), URL https://cran.r-project.org/web/packages/BTYDplus/vignettes/BTYDplus-HowTo.pdf

Herniter, A Probabilistic Market Model of Purchase Timing and Brand Selection (1971), Marketing Science

Platzer & Reutterer, Ticking away the moments: Timing regularity helps to better predict customer activity (2016), Marketing Science