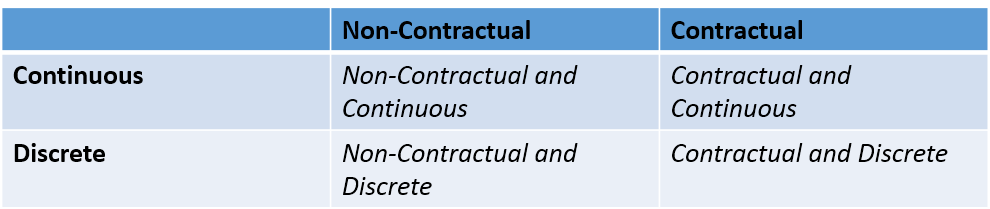
**Customer Lifetime Value Analysis on Retail Data**

*By: Xiaozhou (Joe) Zou*

**I. Problem Analysis**

**1.1 CLV Problem Overview**

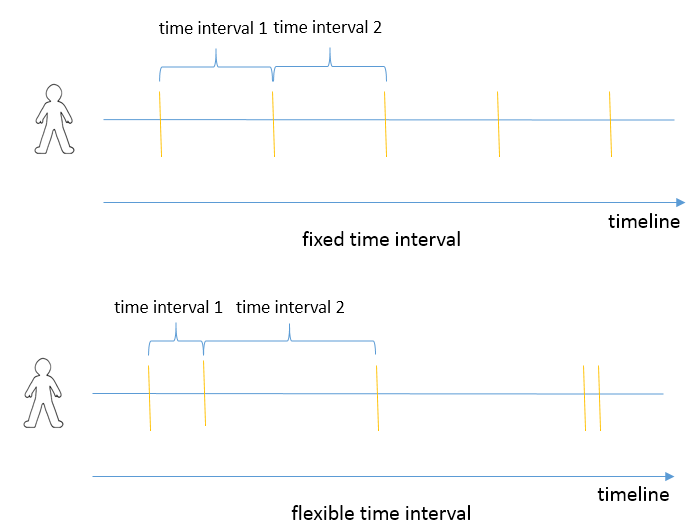
In this homework, my task is to perform a customer lifetime value analysis based on the customer purchase behavior dataset. In particular, I am asked to answer the following two questions: 1. Which 100 customers will statistically make the most purchases in the next 12 months. 2. Which 100 customers will statistically spend the most in the next 12 months. To finish these tasks, I’m going to use lifetime value models. When it comes to customer lifetime value modeling, the first thing to think about is what situation this problem belongs to. Usually there are four situations: continuous V.S. discrete; contractual V.S. non-contractual. The four situations are shown as follows:



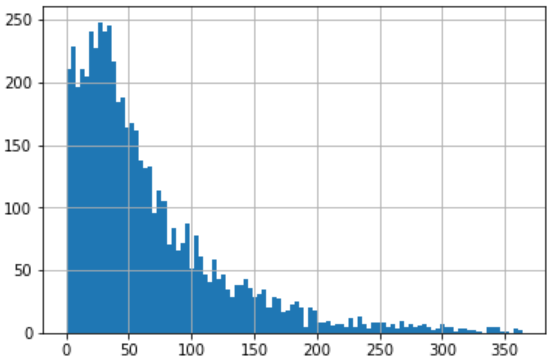
In a discrete setting, purchases occur at fixed frequencies. In the continuous setting, on the contrary, purchases can happen at any time. When there is no contract, customers can leave at any time they want. When there is a contract, however, customers should not leave within the effective period of the contract and if they do, they’ll pay a high cost for that. Once it is known in which scenario this problem falls, there are models ready to be used. So the first task is to make sure in which situation this problem falls.

**1.2 Continuous V.S. Discrete**

At first a decision needs to be made as for whether purchases in this dataset are discrete or continuous. To know whether purchases happen at a fixed frequency or not, time intervals between purchases need to be checked. The following figure compares the situation that corresponds to a fixed purchase interval with that which corresponds to a flexible purchase interval:



When the time interval is fixed, the difference between time intervals are basically 0. On the contrary, when the time interval is not fixed, one time interval differs greatly from another. Based on this observation, the average time interval difference is used to determine whether the purchase behavior in this dataset is a continuous one or discrete one. The average time interval difference is calculated as follows: suppose to are timestamps that correspond to purchases done by customer . Then the purchase time intervals are calculated as , , … , . The difference of purchase time intervals is calculated by subtracting a time interval by one that follows it and then taking the absolute value, i.e., , , …, . Finally, for each customer, the average purchase time interval difference is calculated as . The following is the histogram of the average time interval difference (APTI), which gives a rough description of it.



The histogram gives an overall description of the average purchase time interval. However, it is weak in helping make a decision whether purchases behavior in this dataset corresponds to a continuous one or a discrete one. To make a decision on which category this dataset falls, all customers are divided into two groups. Customers whose APTI is greater than 30 are defined to have a flexible purchase behavior. On the contrary, those whose APTI are smaller than 30 are said to have a fixed purchase behavior. In APTI, the number refers to the number of days, so 30 refers to a month. There is no theorem behind why 30 is selected as the threshold. Since there is no background information pertaining to how instable the purchase behaviors should be to be defined as flexible, this threshold is selected based on my common sense. If in practice, there are more background information that help better decide this threshold, a more accurate decision about whether this purchase behavior is flexible or fixed can be made. In this an exercise, 30 is used as a threshold so that we can move forward to the next step of this solution.

Later the ratio of number of customer in the flexible group to the number of customer in the fixed group is used as *continuity index* to indicate whether the purchase behavior in this dataset is continuous (flexible) or discrete (fixed). In this dataset, the continuity index is 2.064, which means that there are twice as many customers in the flexible group as in the fixed group. So it is reasonable to categorize customers this dataset into ones that have continuous purchase behavior.

**1.3 Contextual V.S. Non-Contextual**

The next step is to decide whether the scenario in this dataset falls into the contractual setting or non-contractual setting. The following analysis assumes the non-contractual setting for two reasons: (1). there is no contract information related in the background information; (2) the business context of this problem, people buying widgets, is more likely to fall in non-contractual situation than in contractual situation. Summarizing the aforementioned, the business in this dataset is likely to be continuous and non-contractual. The following analysis will be based on this point.

**II. Choice of Methodology**

**2.1 Overview of Three Methodologies for CLV Calculation**

When it comes to analyzing customer lifetime value, the most classic method is to use probabilistic models. There are two alternatives that are also often used. The first is to calculate CLV based on some equations. Since there is no modeling process involved in this method, it is more efficient than the probabilistic model. However, it may fail in mining some latent information that could have been captured by a model. The second method to calculate CLV is to use machine learning algorithms. The problem with machine learning model is that it usually takes more time to be well trained in that it often has more parameters. What’s more, when the data contains only the timestamp and monetary value of a purchase, a well-trained machine learning model has similar performance to a probabilistic model. In this data set, there is no additional information except purchase timestamps and monetary value, which makes using probabilistic model is a more time-efficient choice. Summarizing the aforementioned, using probabilistic models is a safe choice for CLV calculation.

**2.2 High Level Description of Probabilistic Models**

Although different probabilistic models make different assumptions on data. All probabilistic models share a similar framework. In this framework, the purchase behavior is characterized by three key parameters:

1. ***Lifetime*:** The period that the customer may purchase the product.

2. ***Purchase rate***: The number of purchases a customer may make in a certain period of time.

3. ***Monetary value***: The expected monetary value of a future transaction.

All probabilistic models assume that those three key latent parameters follow certain probability distributions. The difference among those probabilistic models is that each model has its own assumption of how those three parameters are distributed and how to model the heterogeneity of these three parameters among customers.

**2.3 An example of probabilistic Model: Pareto/NBD model**

For example, Pareto/NBD model, one that is commonly used to model customer purchase behavior in the continuous non-contractual business situation, makes the assumption that the purchase count follows a Poisson distribution with rate **λ;** Lifetime distribution follows and exponential distribution with slope. Each customer has different **λ and**  **that characterize their respective purchase behavior. The difference of λ among customers is modeled by Gamma distribution with parameters ; similarly, the difference ofamong customers is captured by a Gamma distribution parametrized by . The training process of a** Pareto/NBD model is to optimize the four parameters by the data.

**2.4 How to Validate Effectiveness of Model**

**Since the probabilistic models makes certain statistical assumptions on the data, it is desirable that at first those assumptions be validated so that the models are more likely to be effective. Usually such validation is done via hypothesis testing and a resulting p-value is used to show how likely an assumption holds. However, in this problem, cross validation is used instead of hypothesis testing for the following two reasons:**

1. **This is a real-life business problem to which the company has to give an answer. It is possible that for all customer purchase models, the hypothesis testing fails. That way, there is no model left to use. When using cross validation, on the contrast, the optimal model is one that corresponds to the optimal goodness score and always exists. This means that it is ensured that a reasonable answer can be given to the customer.**
2. **The result given by p-value is kind of misleading. For example, suppose hypothesis testing is used to test whether two distributions are the same and the resulting p-value shows that it is almost for sure that they are not the same. However, it may be the case that although the two distributions are not the same, they are actually very similar. That way, the model may still have a pretty good performance but it will be discarded if the model is selected purely by hypothesis testing. Hypothesis testing doesn’t give the distance between two distributions.**

**III. Model Fitting and Validation**

**3.1 Cross-validation Overview**

In this solution, the model is validated via cross-validation. Cross-validation, in my implementation, works as follows: At first the data is split into a training set and a validation set. The model only has access to the training set. Its parameters are trained on the training set. After being trained, it makes predictions on the validation set. Those predictions will be compared with the ground truth in the validation set and the difference between the prediction and true labels is used as a metric to show how well this model works. The model that achieves the optimal goodness score will be used as the final model. That model will be trained on the whole dataset then predict the label in the future. The goodness score can also be used to estimate the predictive ability of the model.

**3.2 Evaluation metric**

**The evaluation metric is a number that indicates the goodness of a model. It is usually defined as some difference between the model prediction and the ground truth in the validation set. In this problem, the average absolute difference is used as the metric. The following is the official definition of average absolute difference:**

**Let be customer purchases in the validation set, the subtitle being the ID of the customer. Let be one of the Customer Purchases Models. Suppose be the predictions made by model corresponding to the customers. Then the average absolute difference of model , denoted by , is defined as:**

**When it comes to measuring the difference, it always comes to my mind whether to use the absolute value of difference or to use the percentage. In this problem, since the data is about retail industry, the absolute value of difference seems to more accurately capture the information in data. For example, suppose a person purchased a laptop last year and three laptops this year. It is more accurate, personally speaking, to describe the change in purchase behavior as buying two more computers than to describe it as a 200% increase in purchasing. However, the choice of using absolute value of difference instead of the percentage of change is quite subjective and should be redeemed when more concrete business context is given. For the current situation, since the only business context information given is that the data comes from the retail industry, it seems more reasonable to use the absolute value of difference in the evaluation metric.**

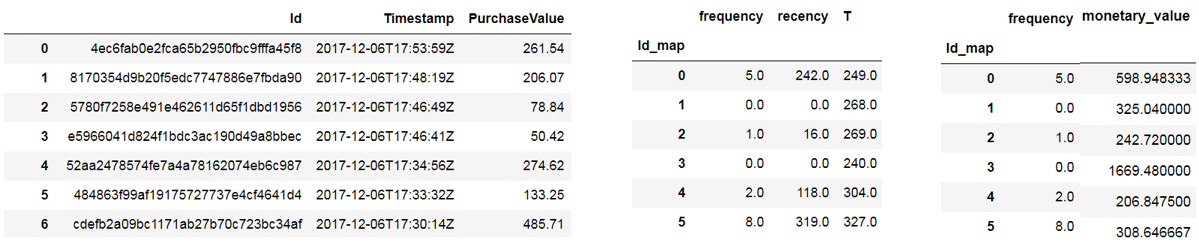
**3.3 Train-Validation-Split**

Based on my research, it is recommended that the time period for the training data should include at least 3 times the typical inter-purchase time of the customers. Since the inter-purchase time is assumed to have an exponential distribution, the “typical” inter-purchase time should be the expectation of the distribution. Using the maximum likelihood estimation, the expectation of the exponential distribution, when estimated from data, is the mean of the sample. In this dataset, the sample mean is 52.231 and there are about one year of purchases. Based on these two statistics, the training period is set to 230 days by trial and error, which is about 4.3 times the sample mean. The rest of the time is in the validation set.

**3.4 Data Pre-processing**

**Before fitting a model, the data should be pre-processed to match the format that the model is suitable for. In the raw data, each record represents a purchase, the three columns represent respectively the id of the customer that performs the purchase, the timestamp of that purchase and monetary value of that purchase. However, the Pareto/NBD model is fit on RFM data, which is the short form for Recency(R), Frequency(F) and Monetary Value(M). The RFM data is indexed by customer ID. Each record contains the information of a customer. By tradition, Frequency is the number of repeat purchases a customer performs. For example, if a customer has in total performed 10 purchases, then the Frequency of that customer is 9, which is the total number of purchases minus 1. The Recency is the time difference between the first purchase and the last purchase of that customer. For example, if the first purchase is done on April 4th and the last on April 20th, then the Recency equals to 16, if measured by days. The Monetary Value is the average monetary value of the purchases done by that customer.**

**The purchase count and monetary value are modelled by two separate models. The Pareto/NBD model, the one for modelling purchase count, requires that each record be in the format of , the meaning of and is defined above and is the complete observation period of that customer, which equals to the time difference between the latest time in this dataset and the time of the first purchase. The following figure shows the snapshot of the raw data (left), the data ready for purchase count modeling (middle) and the data ready for purchase monetary value modeling (right).**

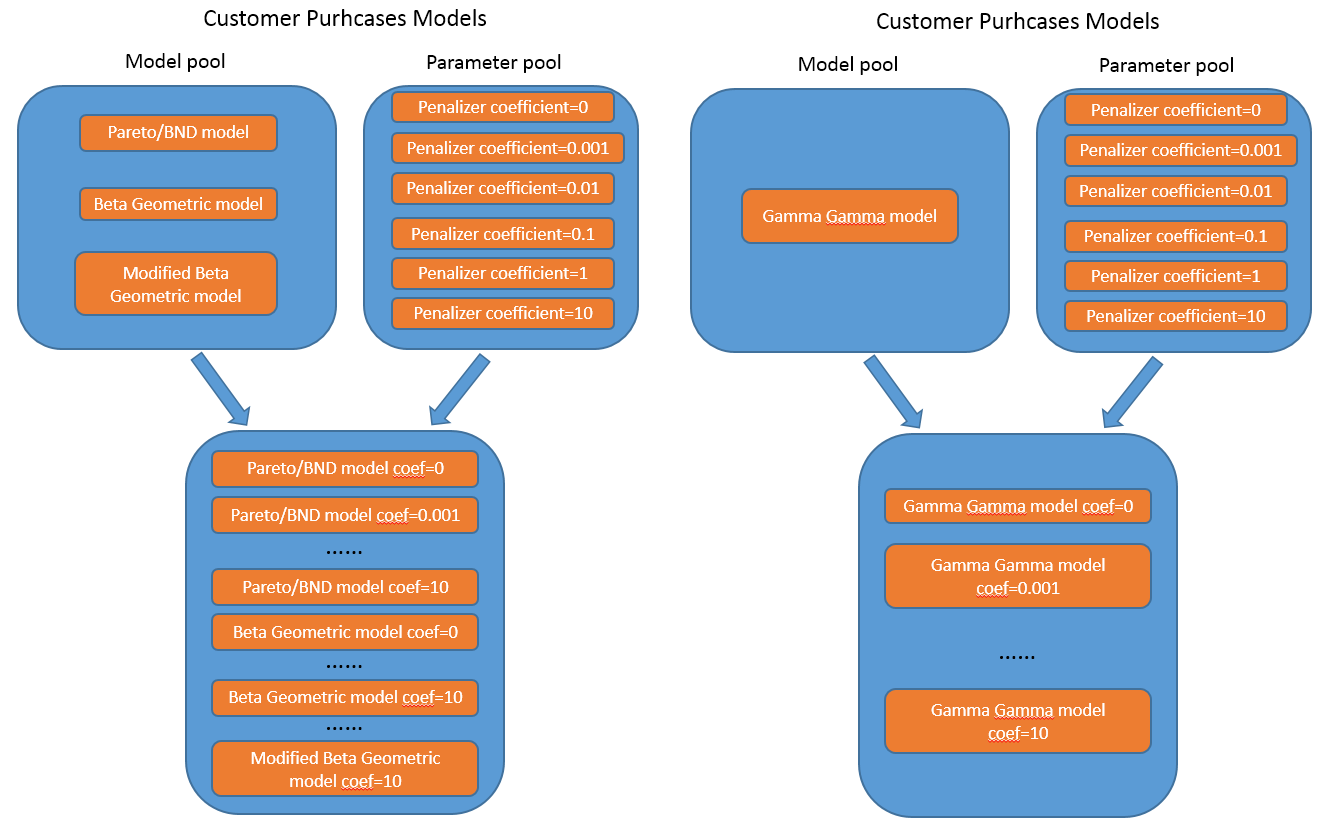


**3.5 Modeling Tool Introduction**

The statistical modeling tool that is use to solve this problem is lifetimes, which is a Python package that’s specifically used to solve the customer lifetime value problem. The greatest advantage of using a domain-specific package is that lots of domain-relevant and technical problems are handled within that package so that concentrations can be focused on the pre-processing the data and building the machine learning workflow.

**3.6 Model Generation and Validation**

The lifetimes package provides four different models (***Pareto/Negative Binomial Distribution model*, *Beta Geometric model, Modified Beta Geometric model*** and ***Beta Geometric Beta Binomial model***) for modeling customer purchases and ***Gamma Gamma model*** for modeling the monetary value of customer purchases. Since the Beta Geometric Beta Binomial model is for discrete business scenario, this model is not taken into consideration in this problem, which has been validated to fall in the continuous scenario. For each of the four models that might be used for modeling and prediction, the package gives control to the user one parameter called penalizer coefficient, which acts as a regularization to prevent the model from overfitting. The optimal penalizer coefficient will be selected by cross validation and candidates are 0, 0.001, 0.01, 0.1, 1 and 10. So for customer purchases modeling, there are in total 18 models to select from (3×6, three refers to the number of different models, six refers to the number of different penalizer coefficients); for monetary value of customer purchase, similarly, there are in total 6 models to select from (1×6, one refers to the number of different model, six refers to the number of different penalizer coefficients). The generation of different models is illustrated below:



**The following is the pseudo-code for the model selection process:**

**Using this model selection algorithm, the optimal model for the Customer Purchase is Beta Geometric Model with penalizer coefficient 0.001; the optimal model for the Monetary Value of Customer Purchase is Gamma Gamma Model with penalizer coefficient 0.**

**IV. Answering the Customer Lifetime Value Business Questions**

**Having selected the optimal model for Customer purchases and the Monetary Value, the next step is to make an inference using these two selected models. When making an inference, the optimal models is trained on the whole data set. The output of Beta Geometric Model is the number of purchases and the output of Gamma Gamma Model is the average monetary value of each customer. To obtain the top 100 customers that spend the most in the next year, those two predictions need to be multiplied. A part of the result is pasted below:**

