Auxiliary Decision-making for Controlled Experiments based on Mid-term Treatment Effect Prediction: Applications in Ant Financial's Offline-payment Business

Gang Li[1] and Huizhi Xie[2]

[1]*Ant Financial Services Group, Internet Financial Center, Haidian, Beijing, China*

[2]*Ant Financial Services Group, Huanglong International Building, Hangzhou, Zhejiang, China*

Keywords: Controlled Experiments, A/B Testing, Mid-term Treatment Effect Prediction, BG/NBD Model, Counting Metrics.

Abstract: Controlled experiments are commonly used in technology companies for product development, algorithm improvement, marketing strategy evaluation, etc. These experiments are usually run for a short period of time to enable fast business/product iteration. Due to the relatively short lifecycle of these experiments, key business metrics that span a longer window cannot be calculated and compared among different variations of these experiments. This is essentially a treatment effect prediction issue. Research in this paper focuses on experiments in the offline-payment business at Ant Financial. Experiments in this area are usually run for one or two weeks, sometimes even shorter, yet the accumulating window of key business metrics such as payment days, payment counts is one month. In this paper, we apply the classic BG/NBD model(Fader et al., 2005) in marketing to predict users payment behavior based on data collected from the relatively short experimentation periods. The predictions are then used to evaluate the impact on the key business metrics. We compare this method with supervised learning methods and direct modelling of treatment effect as a time series. We show the advantage of the proposed method using data collected from plenty of controlled experiments in Ant Financial. The proposed technique has been integrated into Ant Financial experimentation reporting platform, where metrics based on the predictions are one of the auxiliary evaluation criteria in offline-payment experiments.

1 INTRODUCTION

Controlled experimentation has become a hot topic in the last ten years. Major internet companies, including Microsoft (Kohavi et al., 2007), Google (Tang et al., 2010), Facebook (Bakshy et al., 2014), Netflix (Gomez-Uribe and Hunt, 2016), Airbnb (Lee and Shen, 2018) etc. rely heavily on controlled experiments for product improvement and marketing design. Controlled experiments enable agile and fast iteration and thus are indispensable to innovation. Professor Stefan Thomke from Harvard business school once said "If you want to be good at innovation, you have to be good at business experimentation" (Thomke, 2003).

Ant Financial Group provides comprehensive financial services to hundreds of millions of people in China and all over the world. Its flagship mobile application Alipay is one of the most used applications in China. Data-driven decisions are crucial for the success of Ant's business due to its huge scale and the diversity of its businesses. A few years ago, the executive team at Ant Financial realized the importance of experimentation as a key data-driven tool. They started the initiative to build a company-wide experimentation platform in 2016. This initiative has been proven to be a success. Tens of thousands of experiments were run on this platform since its birth. One of the business areas that rely heavily on experimentation is offline-payment. Offline-payment refers to mobile payment in offline scenarios, such as restaurants, supermarkets, malls, etc. In the offline-payment area, we use experimentation heavily for marketing design, offline-payment product improvement, marketing algorithm iteration, and so on.

Metrics are a key element in experimentation. If you cannot measure the performance of your business, you cannot experiment on it. Metric development for experimentation is a non-trivial process and has been discussed in quite a few papers(Kohavi et al.,

Li, G. and Xie, H.
Auxiliary Decision-making for Controlled Experiments based on Mid-term Treatment Effect Prediction: Applications in Ant Financial's Offline-payment Business.
DOI: 10.5220/0009770500190030
In *Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA 2020)*, pages 19-30
ISBN: 978-989-758-440-4

2007; Roy, 2001). Kohavi and coauthors (Kohavi et al., 2007) introduced the so-called overall evaluation criterion (OEC) in their 2007 KDD paper. In addition to the challenges of finding OEC with good directionality and sensitivity, another challenge is to include metrics that reflect an improvement in the long term. Experiments with good short-term performance do not necessarily enjoy good long-term performance (Kohavi et al., 2012). The difficulty of measuring long-term effect comes from the nature of experimentation. Controlled experiments are typically designed to run for a relatively short period of time to enable fast business/product iteration. From the perspective of statistical modelling, correlating short-term effect with long-term effect is essentially a prediction problem. As important for businesses as it is to measure long-term effect, there is relatively little literature on this topic, which reflects the difficulty of this problem. Hohnhold et al. (Hohnhold et al., 2015) developed an experiment methodology for quantifying long-term user learning. They applied the methodology to ads shown on Google search and created a model that uses metrics measurable in the short-term to predict the long-term performance. Dmitriev et al. (Dmitriev et al., 2016) shared pitfalls of long-term online controlled experiments, including cookie instability, survivorship bias, selection bias, perceived trends, side effects, seasonality etc. They also suggested some methodologies to partially address some of these pitfalls. In the area of offline-payment at Ant Finacial, we also suffer from the issue of only observing short-term effect in the experiments. However, our challenge is slightly lower in the sense that we do not need to predict long-term effect but rather mid-term effect. Using the same terminology as in (Hohnhold et al., 2015), long-term effect is what would happen if the experiment launches and users receive the experiment treatment in perpetuity. Mid-term effect refers to impact in the range of a few weeks to a few months. The business argument for focusing on mid-term effect in offline-payment is that it is a relatively new (compared to areas such as search) area and things change much faster.

In this paper, we focus on the prediction of two OEC metrics in offline-payment: payment days and payment counts. Both are counting metrics. Their definition is given in Section 2.1 and also Section 4. We propose to use stochastic process models for counting metrics to tackle the prediction problem. The main advantage of these stochastic process models are that we do not need to train the models based on pre-experiment data. The models are trained using data collected in the experimentation period, separately for control and treatment, and thus potentially have higher prediction accuracy. We first review a few well-known stochastic process models in the marketing area, including Pareto/NBD, BG/NBD, and their extensions. Although these models were developed in the marketing area, they can potentially be used to model any counting metric. We then apply these models in the context of experimentation. We define three levels of accuracy in model evaluation. They are metric prediction accuracy, treatment effect prediction accuracy, and decision-making prediction accuracy. We explicitly call out these three levels of accuracy and show that high decision-making prediction accuracy is much easier to achieve than the other two. Hence it is meaningful to invest in prediction in the context of experimentation. We demonstrate the effectiveness of this methodology using experiments in the area of offline-payment at Ant Financial. The proposed methodology has been integrated into Ant Financial's experimentation reporting platform, where metrics based on the predictions are one of the key evaluation criteria in offline-payment experiments.

The contributions of this paper are summarized as follows. i) To the best of our knowledge, we are the first to apply stochastic process models for treatment effect prediction in controlled experiments. These stochastic process models are built separately for control and treatment versions in each experiment on the fly. Hence they enjoy higher prediction accuracy for both control and treatment versions, and finally treatment effect. ii) We propose three levels of prediction accuracy: metric level, treatment effect level, and decision-making level. We show that decision-making level accuracy is most attainable and also most meaningful from the decision-making point of view. iii) We provide numerous case studies based on real data from Ant Financial and show the effectiveness of the proposed solution in this paper.

The remainder of this paper is organized as follows. We give a brief overview of controlled experiments in the offline-payment business at Ant Financial in Section 2.1. A few well-known stochastic process models for counting metrics are reviewed in Section 2.2. The application of stochastic process models to predict mid-term treatment effect is discussed in Section 3. In Section 4, we share case studies at Ant Financial to demonstrate the effectiveness of the idea in this paper. We conclude the paper in Section 5, where we summarize the work in this paper as well as a few future research directions.

2 BACKGROUND AND OVERVIEW OF STOCHASTIC PROCESS MODELS FOR COUNTING METRICS

2.1 Brief Overview of Controlled Experiments in Offline-payment at Ant Financial

As mentioned in the introduction section, most changes in the offline-payment business at Ant Financial are evaluated via controlled experiments prior to launch. These changes include marketing algorithm iteration, marketing strategy development, payment product improvement, and so on. Due to the page limit, we do not review the basic concepts of controlled experiments in this paper. Readers can refer to (Kohavi et al., 2007) for details. The two key business metrics in offline-payment are payment days and payment counts in the window of a few months. Payment days is calculated as follows. Given a time period, number of days with at least one payment for each user is calculated, then a sum is taken across all users as the payment days metric. Similarly for payment counts, we first calculate number of payments for each user and then take a sum. Payment days and payment counts are essentially counting metrics. They are first calculated at user level and then the statistical summary sum is calculated. In controlled experiments, since the traffic percentage between control and treatment are not necessarily the same, we typically use average instead of sum as the statistical summary. Statistical inference for the comparison between control and treatment is conducted using standard two-sample t-test thanks to the large sample sizes.

2.2 Stochastic Process Models for Counting Metrics

Stochastic process models for counting metrics in marketing can be classified into two categories: noncontractual scenarios and contractual scenarios. Typical examples of contractual scenarios include cell phone services, bank services, etc (Fader and Hardie, 2007). Customer's relationship with Ant Financial in offline payment is noncontractual. We therefore review a few well-known stochastic process models in the noncontractual scenario and their recent applications (Dahana et al., 2019; Dechant et al., 2019; Venkatesan et al., 2019).

2.2.1 The Pareto/NBD Model

The Pareto/NBD model was proposed by Schmittlein et al. in (Schmittlein et al., 1987) to model repetitive purchase behavior. Under the model, customers drop out with a certain probability at any given time. The dropout is unobservable due to the noncontractual nature. For customer i, define the following notations.

- x_i is the number of purchases made by this customer in $(0,T_i]$, where $(0,T_i]$ is the observation window for this customer. Note that customers come into observation at different times and thus the observation window varies across customers. The starting point "0" is the time of the first purchase. The calculation x_i excludes the first purchase.
- t_{x_i} is the time of the last purchase in the observation window.
- T_i is the observation length of customer i. T_i also varies across customers.

The main assumptions of the Pareto/NBD model are as follows.

1. For a given active customer, the repetitive purchase behavior of this customer follows a Poisson process with transaction rate λ_i. Active customers refer to those that have not dropped out.
2. For any given customer, let τ_i denote the life time of this customer. Note that τ_i is not observable in the noncontractual scenario. τ_i follows an exponential distribution with dropout rate μ_i.
3. λ_i varies across customers and follows a gamma distribution with parameters (r,α).
4. μ_i varies across customers and follows a gamma distribution with parameters (s,β).
5. λ_i and μ_i are independent from each other.

The Poisson process repetitive purchase behavior and exponential life time imply the lack of memory property. These assumptions have been proven to hold in many marketing scenarios (Ehrenberg, 1972; Karlin, 2014). We will discuss the validation of these assumptions in the case study section.

For a fixed observation window, the input of the Pareto/NBD model are the tuples (x_i,t_{x_i},T_i) of customers. The parameters (r,α,s,β) are estimated using the maximum likelihood estimation technique. For any given customer i, the two important output of the Pareto/NBD model are as follows.

- $E(Y_i(t)|X_i = x_i,t_{x_i},T_i,r,\alpha,s,\beta)$: the conditional expectation of number of purchases in a future period $(T_i,T_i+t]$

- $P(\tau_i > T_i | X_i = x_i, t_{x_i}, T_i, r, \alpha, s, \beta)$: the conditional probability of dropping out after T_i or the conditional probability of being active after time T_i

These two outputs can be used to calculate most of the common managerial questions such as:

- How many (active) retail customers does the firm now have?
- Which individuals on this list most likely represent active customers? Inactive customers?
- What level of transactions (for example offline-payment counts) should be expected next month by those on the list, both individually and collectively?

2.2.2 The BG/NBD Model

Despite the solid theoretical foundation of the Pareto/NBD model, people have found it hard to use because of the efforts needed to estimate the parameters. Recall one of the key assumptions of the Pareto/NBD model is that customers can drop out at any given time, independent of their purchase behavior. This assumption implies that the dropout process is continuous, which makes the optimization of the likelihood function difficult. To solve this issue, Fader et al. developed the beta-geometric/NBD (BG/NBD) model in 2005 ((Fader et al., 2005)). The key difference of the BG/NBD model from the Pareto/NBD model is that it assumes the dropout of a customer can only occur immediately after a purchase. This slight variation makes the BG/NBD model much easier to implement. Interested readers can refer to (Fader et al., 2005) for more details. The BG/NBD model has been proven to have similar performance as the Pareto/NBD model in terms of prediction accuracy in many applications (Trinh, 2013; Dziurzynski et al., 2014).

The input of the BG/NBD model is exactly the same as that of the Pareto/NBD model. The assumptions for BG/NBD model is listed as follows:

1. For a given active customer, the repetitive purchase behavior of this customer follows a Poisson process with transaction rate λ_i.
2. An active customer drops out with probability p_i after a purchase. Therefore, the total number of purchases J_i of a customer before dropout follows a geometric distribution $P(J_i = j | p_i) = p_i(1-p_i)^{j-1}$.
3. λ_i varies across customers and follows a gamma distribution with parameters (r, α)
4. p_i varies across customers and follows a beta distribution with parameters (a, b).
5. λ_i and p_i are independent from each other.

Parameter estimation of BG/NBD is much easier than Pareto/NBD model, and outputs of the BG/NBD model is similar to that of the Pareto/NBD model. For any given customer i, the two main output are as follows.

- $E(Y_i(t) | X_i = x_i, t_{x_i}, T_i, r, \alpha, a, b)$ (Fader et al., 2005)
- $P(\tau_i > T_i | X_i = x_i, t_{x_i}, T_i, r, \alpha, a, b)$ (Fader et al., 2008)

Both Pareto/NBD and BG/NBD model purchases in continuous time, i.e., purchases can happen at any time. However, some businesses track repeat purchases on a discrete-time basis. To model purchases on a discrete-time basis, Fader et al. proposed the discrete-time analog of the BG/NBD model, the beta-geometric/beta-Bernoulli (BG/BB) model. The details of this model can be found in the original paper(Fader et al., 2010) and omitted here due to page limit.

2.2.3 A Hierarchical Bayes Extension to the Pareto/NBD Model

In the Pareto/NBD and BG/NBD models, the heterogeneity of transaction rate and dropout rate across customers are modelled using a single distribution separately. Individual-level rates cannot be estimated. Also, the independence assumption of these two random variables is hard if not impossible to verify. In order to address these issues, Abe extended the Pareto/NBD model using a hierarchical Bayesian framework in (Abe, 2009). The hierarchical Bayesian extension allows incorporation of customer characteristics as covariates, which can potentially increase prediction accuracy and also relax the independence assumption. The Hierarchical Bayes Extension (HBE) model is based on the following assumptions.

1. A customer's relationship with the merchant has two phases: alive and dead. This customer's lifetime τ_i is unobserved and follows an exponential distribution with dropout rate μ_i, i.e. $f(\tau_i | \mu_i) = \mu_i e^{-\mu_i \tau_i}$
2. While alive, this customer purchase behavior follows a Poisson process with transaction rate λ_i, i.e. $P(x_i | \lambda_i, t_{x_i}) = \frac{(\lambda_i t_{x_i})^{x_i}}{x_i!} e^{-\lambda_i t_{x_i}}, t_{x_i} \leq T_i$
3. λ and μ follow a bivariate lognormal distribution, i.e. $\begin{bmatrix} \log(\lambda_i) \\ \log(\mu_i) \end{bmatrix} \sim \mathbf{BVN}(\theta_0 = \begin{bmatrix} \theta_\lambda \\ \theta_\mu \end{bmatrix}, \Gamma_0 = \begin{bmatrix} \sigma_\lambda^2 & \sigma_{\lambda\mu} \\ \sigma_{\lambda\mu} & \sigma_\mu^2 \end{bmatrix})$ where BVN denotes the bivariate normal distribution.

4. λ and μ are correlated with some covariates such as customer characteristics through a linear regression model, i.e.
$\begin{bmatrix} \log(\lambda_i) \\ \log(\mu_i) \end{bmatrix} = \beta^T d_i + e_i$
where $\beta \in \mathbb{R}^{k\times 2}$ is the regression coefficients vector, $d_i \in \mathbb{R}^k$ is the covariate vector and $e_i \sim \mathbf{BVN}(0, \Gamma_0)$ is random error.

The first two assumptions of the HBE model are exactly the same as those of the Pareto/NBD model. The lognormal and linear model assumptions are mainly for mathematical convenience. The input to the HBE model includes customer's transaction information and covariate information. The model parameters are estimated using the MCMC procedure. For more details, please refer to (Abe, 2009). Since covariate information is included, the HBE model can produce individual-level transaction rate and dropout rate estimates. The prediction outputs are the same as the aforementioned models and thus not repeated.

3 PREDICTING MID-TERM TREATMENT EFFECT IN CONTROLLED EXPERIMENTS

In this section, we discuss the application of the aforementioned stochastic process models in controlled experiments. For a given metric, treatment effect is defined as the expected difference of this metric between control and treatment. In the offline-payment business, the two OEC metrics are payment counts and payment days over one month. The detailed calculation was presented in the background section and thus not repeated here. The observation window of metrics in controlled experiments is the experimentation period. Customers enter an experiment at different times. Hence the observation length of each customer is different. The dashboard of experiment results showing treatment effects is typically updated on a daily basis. After an experiment starts, on a given day, the observation window is from the beginning of the experiment to the given day. For customer i, the time of the first purchase during this observation period is set as "0", and T_i is the duration between time "0" and the given day (for continuous time, we use the last second of the given day). we collect the following transaction statistics in $(0, T_i]$.

- number of **repeated** payments or number of **repeated** days with payment, denoted as x_i. Note that the number of payment is for the metric payment counts and the number of days with payment is for the metric payment days. Note that the first purchase is not included.
- the latest payment time (day) t_{x_i}: if $x_i > 0$ then $t_{x_i} > 0$ else $t_{x_i} = 0$
- observation length T_i: the time difference between the first purchase and the given day (ending point of observation period)

Since the dashboard of experiment results is updated on a daily basis, these models are also retrained on a daily basis using the above collected information as input, separately for control and treatment. With the trained models, we can then predict individual-level number of payments and number of days with payment in a future period. These predictions are treated as user-level predictions, and we predict the OEC metrics by averaging these predictions across customers. The prediction of treatment effect refers to the calculation of treatment effect based on the metric predictions. Specifically, we define the prediction of treatment effect as the difference of metrics between control and treatment. Statistical inference of these predictions is done similarly as for metrics observed in the experimentation period.

The accuracy of the predictions in the context of controlled experiments is evaluated at three levels, from the most difficult to the easiest.

- The first level is metric level prediction accuracy. It is defined as the difference in the average of the actual metric value and the predictions for a cohort of users, e.g. users in control.
- The second level is treatment effect level prediction accuracy. It is defined as the difference in the observed treatment effect and the predicted treatment effect.
- The third level is decision-making level prediction accuracy. At the stage of decision-making, there are three possible outcomes: control is statistically significantly better than treatment; control is statistically significantly worse than treatment; control and treatment are not statistically significantly different. Decision-making level prediction accuracy is defined as the proportion of decisions where statistical inference based on observed metrics agree with that based on predicted metrics.

In general, it may be easier to achieve high prediction accuracy at treatment effect level than metric level. This is because prediction error for control and treatment can be biased toward the same direction and thus cancel each other. Decision-making level prediction is easier than treatment effect level because it essentially looks at the sign, not the actual value.

There are two things worth noting here. First, for a given observation period, users without purchases do not participate in model training. The predictions for them in a given future period are 0. Note this is by design and can be a future improvement direction. Secondly, as will be seen in the case study section, variance of the predicted metrics is usually lower than the observed metrics. The possible explanation is as follows. Let M and $\hat{M}$ denote the actual metric and the predicted metric respectively. We have $M = \hat{M} + e$, where e is the variation in M that is not accounted for by $\hat{M}$. If we can assume the independence between e and M (in linear models we can show that this is true), then it is trivial to show that $Var(M) > Var(\hat{M})$.

4 CASE STUDIES AT ANT FINANCIAL

In this section, we show some numerical results of the application of the aforementioned models to controlled experiments in the offline-payment business at Ant Financial. Again, the two metrics of interest are payment counts and payment days. Real values are masked for the purpose of data security. Throughout this section, **APC** and **APD** refer to the average payment counts and the average payment days across customers respectively. More precisely, **APC** $= \frac{\Sigma_i \Sigma_j \text{paycount}_{ij}}{\Sigma i}$ and **APD** $= \frac{\Sigma_i \Sigma_j \text{payday}_{ij}}{\Sigma i}$ where paycount_{ij} refers to the paycounts for customer i in day j and $\text{payday}_{ij} = 1$ if $\text{paycount}_{ij} > 0$ else 0. There are several remarks regarding the data used in the paper as follows:

1. The data used in this research does not involve any Personal Identifiable Information (PII)
2. The data used in this research were all processed by data abstraction and data encryption, and the researchers were unable to restore the original data.
3. Sufficient data protection was carried out during the process of experiments to prevent the data leakage and the data was destroyed after the experiments were finished.
4. The data is only used for academic research and sampled from the original data, therefore it does not represent any real business situation in Ant Financial Services Group.

4.1 Results of the BG/NBD Model

We start with the BG/NBD model since it is easier to implement and has been shown to have similar performance as the Pareto/NBD model. The first example is based on an algorithm experiment. Number of users

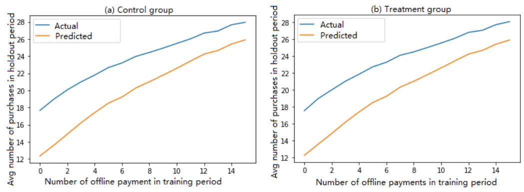

Figure 1: Metric Level Prediction Accuracy of APC by User Cohort in the Holdout Period for (a) Control and (b) Treatment in Example 1.

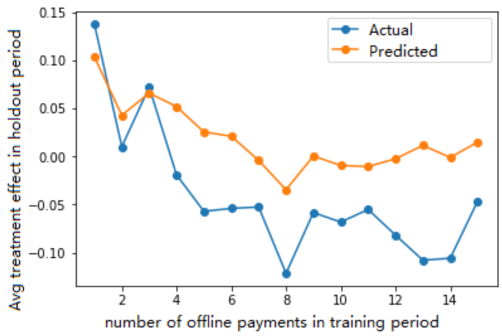

Figure 2: Treatment Effect Level Prediction Accuracy of APC by User Cohort in the Holdout Period in Example 1.

in control and treatment are both at the level of tens of thousands. The training period is the first two weeks of the experiment period. The holdout or prediction period is the 30-day window right after the first two weeks. We present metric level prediction accuracy with respect to number of purchases in the training period. The results are in Figure 1 for control and treatment separately. There is significant bias in metric level prediction accuracy. The bias can be due to either internal operational activities such as new promotions or external changes such as promotions from competitors. This proves the difficulty of metric level prediction. The bias seems to decrease for heavy users. This is probably due to their higher contribution in terms of data volume in the training period. Treatment effect level prediction is presented in Figure 2. The absolute bias is much smaller because bias in control and treatment to some extent cancels each other. Decision-making level results are presented in Table 1. The prediction result agrees with the actual outcome.

To be more convincing, we share results of a few more examples. The setting of these examples and prediction results are presented in Table 2 and Table 3 respectively. We have the following observations from Table 3.

- It is very difficult to achieve high metric level prediction accuracy. This is expected. Payment counts and payment days can both be affected by many factors, within or outside Ant Financial.

Table 1: Prediction Accuracy in the Holdout Period for Example 1.

	Observed APC	Predicted APC	Observed APD	Predicted APD
Treatment Control	33.55 33.45	28.95 28.75	10.8 10.75	8.6 8.5
Treatment Effects (std)	0.1 (0.02)	0.2 (0.02)	0.05 (0.007)	0.1 (0.005)
Inference Results	Sig. higher	Sig. higher	Sig. higher	Sig. higher

'Sig. higher' means 'statistically significantly higher'. 'std' refers to 'standard deviation'.

Table 2: Application Setting of Examples 2-5.

	Number of Users	Training Duration	Holdout Duration
Example 2	Tens of thousands	6 days	8 days
Example 3	Millions	8 days	29 days
Example 4	Millions	6 days	11 days
Example 5	Millions	30 days	30 days

For example, a promotion event or a holiday can change the distribution of these metrics significantly. We will provide a more specific example shortly.

- The standard deviation of predicted treatment effect is indeed lower than that of the observed treatment effect. The intuition was briefly discussed in Section 3. The results here empirically confirmed the intuition.
- The absolute difference between predicted treatment effect and observed treatment effect is usually much smaller than that between predicted metric value and observed metric value. However, the relative difference in treatment effect is not necessarily smaller.
- Decision-making level prediction accuracy is significantly higher than the other two. In Section 4.4, we will show more results of decision-making level prediction accuracy.

In Figure 3, we show the temporal trend of the prediction accuracy in the holdout period of Example 3. The x-axis is the number of days since the beginning of the holdout period. Note the sudden drop of prediction accuracy on day 15, which turns out to be the Lunar Spring Festival. Although the holiday causes significant bias in the prediction in both control and treatment (can be seen as the seasonality of prediction task), the treatment effect prediction accuracy as shown in Figure 4 seems to be much less affected. This is another typical example where the prediction bias in control and treatment cancel each other.

Table 3: Prediction Accuracy in the Holdout Period for Examples 2-5.

	Obs APC	Pred APC	Obs APD	Pred APD
Treatment2 Control 2 TE 2 (std) IR 2	6.26 6.36 -0.1 (0.054) Not sig.	7.09 7.25 -0.16 (0.05) Sig. lower	2.34 2.37 -0.028 (0.012) Sig. lower	1.91 1.95 -0.03 (0.005) Sig. lower
Treatment3 Control 3 TE 3 (std) IR 3	4.93 4.90 0.032 (0.008) Sig. higher	6.72 6.69 0.035 (0.008) Sig. higher	3.01 2.98 0.035 (0.0034) Sig. higher	3.81 3.79 0.02 (0.0025) Sig. higher
Treatment4 Control 4 TE 4 (std) IR 4	1.71 1.72 -0.009 (0.005) Not sig.	1.78 1.78 -0.004 (0.004) Not sig.	0.86 0.86 -0.001 (0.0016) Not sig.	0.805 0.805 -0.0014 (0.0010) Not sig.
Treatment5 Control 5 TE 5 (std) IR 5	4.63 4.59 0.039 (0.02) Sig. higher	3.09 3.07 0.026 (0.009) Sig. higher	2.73 2.71 0.028 (0.006) Sig. higher	2.02 2.01 0.017 (0.004) Sig. higher

'Sig. higher' means 'statistically significantly higher'.'Sig. lower' means 'statistically significantly lower'. 'Not sig' means 'Not statistically significantly different'. 'std' refers to 'standard deviation'. 'TE' is 'Treatment Effect', 'IR' is 'Inference Results', 'Obs' means 'observed', 'Pred' means 'Predicted'

4.2 Comparison between the BG/NBD Model and Other Models

In this section, we compare the BG/NBD model with other models. We first compared the BG/NBD model with the BG/BB model on metric APD. We expect the BG/BB model to perform better than the BG/NBD model on APD since APD is discrete-time based. However, we did not find any gain in the BG/BB model after numerous example evaluations. The results are not included here due to page limit. We suspect the reason may be the number of transaction opportunities in controlled experiments is not high enough to fully explore the strength of the BG/BB model.

We then compared the BG/NBD model with the hierarchical Bayesian extension of the Pareto/NBD model (called HBE for short in discussion that follows). 40+ user characteristic features that are se-

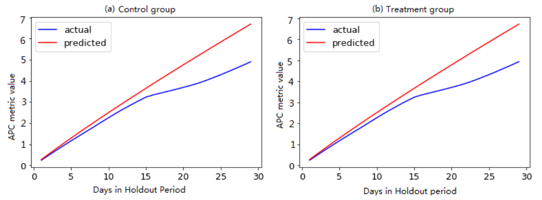

Figure 3: Temporal Trend of APC Prediction for (a) Control Group and (b) Treatment Group in Example 3.

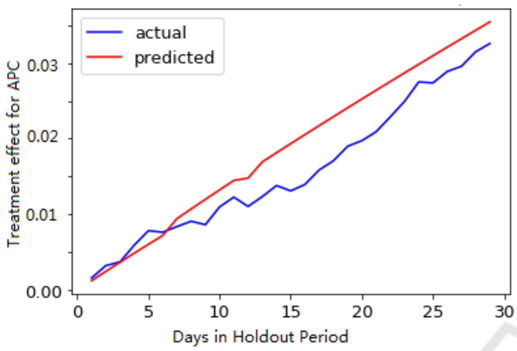

Figure 4: Temporal Trend of APC Treatment Effect Prediction in Example 3.

Table 4: Comparison of BG/NBD and HBE on Prediction Accuracy.

	Model	R^2	MSE
Example 6	BG/NBD	0.438	18
	HBE	0.441	17.93
Example 7	BG/NBD	0.46	59
	HBE	0.41	65

lected from a tree based feature importance procedure are included in the HBE model. The two evaluation criterion are R^2 and mean squared error (MSE) in the holdout period. Since the two criterion are well-known in the literature, we do not repeat the definition of them here. The implementation of the HBE model is quite complicated. The R package "BTYDPlus" is very slow. We had to implement the HBE model in python with some modifications. The key modifications include replacement of the Bayesian regression model with the elastic net model, removing the capping restriction of dropout rate, modification of the sampling window, etc. The implementation details are not included due to page limitation. Results of two out of numerous examples we tried are shown in Table 4. Based on the examples in table 4 and more examples not listed, the performance of the HBE model is not stably outperformed. We could not conclude uniformly better prediction accuracy of the HBE model. In fact, in a non-trivial number of trials, the HBE model performs worse than the BG/NBD model, even after a fine tuning of the hyper-parameters. This may have something to do with the relatively short period of training data or that the user characteristic information is already fully captured in the user purchase behavior. The instability and the high latency of the HBE model make it unsuitable for reporting purpose in controlled experiments.

4.3 Validation for Model Assumptions

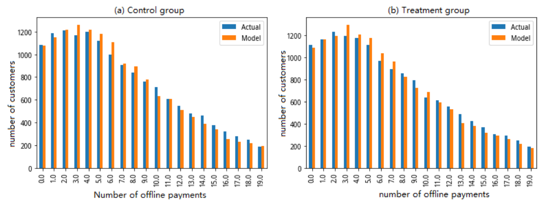

Figure 5: Distribution of Number of Purchases in the Training Period for (a) Control Group and (b)Treatment Group in Example 2.

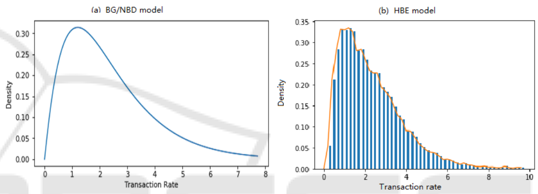

Figure 6: Distribution of Transaction Rate from (a) BG/NBD Model and (b) HBE Model in Example 6.

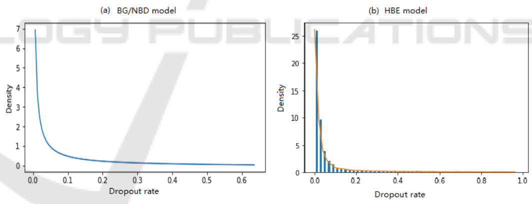

Figure 7: Distribution of Dropout Rate from (a) BG/NBD Model and (b) HBE Model in Example 6.

We discuss the validation for the assumptions of the BG/NBD model in this section. Although it has been proven to work in the past, it is worth checking with our data. The set of assumptions can be decomposed into two parts: (1) Poisson repetitive purchase and exponential life; (2) the Gamma distribution of transaction rate and Beta distribution of dropout rate as well as their independence. For the first part, we compare fitted histogram of number of purchases with the actual and use their closeness as an indirect way to verify the assumption. Note that this comparison is done in the training period and a good fit in the training period does not necessarily lead to a high metric level prediction accuracy in the holdout period. We were able to empirically confirm the validity of part (1) based on quite a few experiments. We share one such example in Figure 5, where the Kullback - Leibler (KL) distance from actual distribution to predicted distribution are 0.0014 and 0.0021 for control and treatment group respectively. KL distance, defined as $D_{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$ where P, Q are two discrete distributions, also called relative entropy , is often used to measure the closeness of two distribution. KL distance has two properties, i) $D_{KL} \in [0, \infty)$, and $D_{KL} = 0$ only when two distributions totally match, ii) the lower KL is , the closeness is higher between two distributions. The above results of KL distance are considerably lower which implies the good fitness of the used model.

For part (2), we rely on results from the HBE model since it can produce user level transaction rate and dropout rate. In the experiment we analyzed, we were also able to confirm this part of the model assumptions. We share one such example in this paper. The distributions of transaction rate and dropout rate based on the BG/NBD model and the HBE model are presented in Figures 6 and 7 respectively. KL distance from Gamma distribution (Fig6a) to estimated transaction rate distribution (Fig6b) is calculated as 0.053, and KL distance from Beta distribution (Fig7a) to estimated dropout rate distribution (Fig7b) is 0.167. The KL distances are small, which indicates a good match for both transaction rate and drop rate. Also, the empirical correlation coefficient between user-level transaction rate and dropout rate is 0.051, which shows the independence assumption is reasonable.

An important thing to note is that the purpose of model assumption verification is really to discover improvement opportunities and future research direction. At the end of the day, what we care about is the decision-making level prediction accuracy.

4.4 Production Results

4.4.1 Baseline Models

Besides stochastic process models, there are other ways to model and predict treatment effect in controlled experiments. We introduce a few such models in this section. The first is to model treatment effect as a time series. Commonly used time series models include ARMA, AR, MA models etc. The main difficulty of using time series models is the lack of data points since controlled experiments are run for only weeks if not days. The second is to model treatment effect as a function of time point t. Let y_t denote the treatment effect at time t. The model can be written as $y_t = f(t) + \varepsilon_t$, where ε_t is random error. Again, due to the lack of data points, $f(t)$ has to be simple. Two such functions are linear function $f(t) = at + b$ and exponential function $f(t) = be^{at}$. Based on the empirical results, the linear function seems to be more robust than the exponential function. Hence we share two examples of the comparison between the BG/NBD model and the linear model. In the two examples, the first seven days is the training period and the BG/NBD model is trained based on data in this period. For the linear model, we treat users in the training period as the cohort of interest and use data from the training period and the first five days of the holdout period to train the linear function. The results are presented in Figures 8 and 9, which clear indicates the linear trend model heavily relies on the fitted trend of early training points. If the trend of training points is consistent with test points, the model works well as shown in Fig.8. If the trend of training points happens to have a large disturbance, the predicted trend deviates from actual trend a lot as shown in Fig. 9. This should be due to the fact that BG/NBD models user-level data while the linear model is trained on summarized data where significant amount of information can be lost.

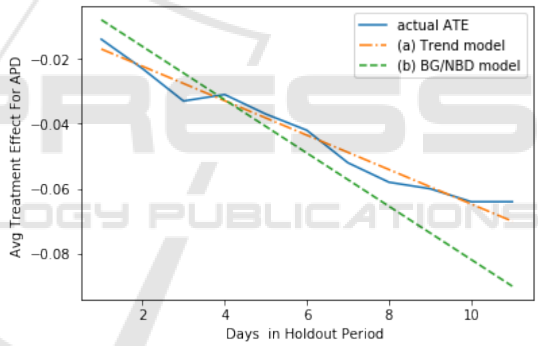

Figure 8: Temporal Trend Prediction of treatment effect in APD with (a) Trend model (b) BG/NBD model in Example 2.

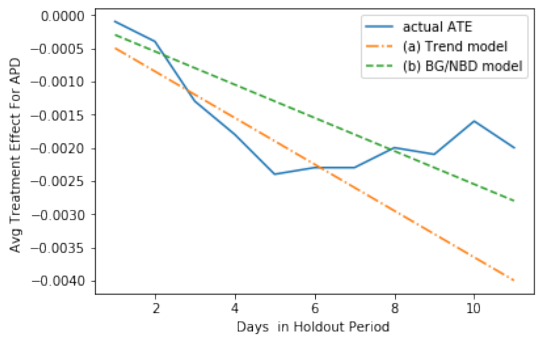

Figure 9: Temporal Trend Prediction of treatment effect in APD with (a) Trend model (b) BG/NBD model in Example 3.

Another baseline model is a supervised learning model. In the supervised learning model, we model user-level data. At a given time, for each user, the target variable is the number of days with payment and number of payments in the next 30 days. Features include user's demographics and behavior data up to the given time. The algorithm we pick is xgboost based on its superior empirical performance compared to other supervised learning models. As mentioned before, for a given experiment, we cannot train the supervised learning model based on data collected in the experiment. This is because there is no label for the target variable. Instead we have to train the model before the experiment based on historical data and then predict for the experiment. Intuitively, change in the distribution of users in the experiment, especially in the treatment variation, can potentially lead to low decision-making level prediction accuracy. We will compare the xgboost model with the BG/NBD model in production.

4.4.2 Comparison Results

Based on the empirical results, we concluded that the BG/NBD model is the most suitable to be productionized for both OEC metrics. We implemented the model using in-house tools and integrated with the experimentation reporting pipeline. As mentioned before, the predictions are treated as metrics and statistical inference is done on them. We show results of thirty-day prediction based on both the BG/NBD model and the aforementioned xgboost model in this section. Decision-making accuracy for thirty-day prediction is presented in Table 5 , which contains 35 experiments or 796 records (experiment version * days) in total. Since the results are based on a limited number of experiments, we also report the 95% confidence interval of the accuracy to incorporate uncertainty of the results with bootstrap method. Note that there are three possible inference outcomes: treatment significantly higher than control; treatment significantly lower than control; treatment not significantly different from control. Hence a random guess of the inference result in a future period would give an accuracy of 33%. We have the following observations from Table 5.

- Prediction accuracy based on the BG/NBD model is much higher than random guess and the xgboost model, which shows the effectiveness of the proposed approach. Even with a 0.78 R^2 in the training data, the prediction accuracy of the xgboost model is very low and even lower than random guess for APD, which is most probably because the model is not trained based on data in the current experiments. Many things can change between the experiments and historical data, e.g. distribution of user characteristics, the relationship between the features and the target variable, etc.
- Prediction accuracy for APC is higher than that for APD, especially when the training duration is short. This is because APD is discrete-time based. The same training duration yield much less information for APD (measured by number of days) than for APC (measured by continuous time). A related observation is that as the duration increases, the gain in prediction accuracy for APD is much more significant than that for APC. In fact, from the 95% confidence intervals, we can see that the prediction accuracy for APC is not significantly different at 5% significance level between the "training duration $<$ 15 days" scenario and the "training duration $\geq$ 15 days" scenario (break-point 15days is set in means of the prediction accuracy for BG/NBD).

Since there is a gap between predicted value and real value for the mid-term OEC metric, and the predicted metric usually has smaller variance than real metric, we suggest to use these predicted mid-term metrics as auxiliary metrics, which tends to reflect the developing trend from the behavior data during the experiment period.

5 CONCLUSIONS

In this paper, we tackle the problem of prediction for counting metrics, with applications in controlled experiments in the offline-payment business at Ant Financial. We propose to use stochastic process models for the prediction purpose. The main advantage of these stochastic process models is that they can be (re)trained on data collected from users in live experiments. Since the training and prediction are done for the same users for different experiment variations separately, the prediction accuracy is potentially higher. The relationship between users and Ant Financial is noncontractual. We thus review and apply well-known stochastic process models in the noncontractual scenario in marketing. With these stochastic process models from the noncontractual scenario, we also do not run into the difficulty of labelling (we do not observe when a user drops out) as in supervised learning models. We empirically compare the models based on data from real experiments in Ant Financial. Based on the empirical results, we conclude that the BG/NBD model is the most suitable for the purpose of mid-term treatment effect prediction in controlled

Table 5: Thirty-day Decision-making Prediction Accuracy in Production.

Model	Stochastic Process model		Regression Model with Xgboost	
	APC	APD	APC	APD
Overall	78.1% [75.1%, 80.8%]	58.4% [54.9%, 61.8%]	39.5% [36.1%, 42.9%]	23.7% [20.1%, 25.9%]
duration $<$ 15 days	79.2% [75.1%, 83.2%]	40.2% [35.3%, 45.0%]	40% [35.1%, 44.8%]	20% [16.0%, 23.9%]
duration $\geq$ 15 days	77.1% [73.0%, 81.1%]	75.8% [71.6%, 79.9%]	39.1% [34.2%, 43.7%]	27% [22.6%, 31.3%]

experiments in the offline-payment business at Ant Financial. Possible explanation of the results is also given. The BG/NBD model has been productionized. Production results show the effectiveness of the proposed methodology. Analysis of the effect of training duration on prediction accuracy is also conducted. This analysis is very useful to guide the operation of controlled experiments, e.g., decide the run time of experiments. Two possible future directions are as follows. First, although the stochastic process models are from the marketing area, they can be used to model any counting metric. Hence extension to other counting metrics in Ant Financial is desired. The second direction is to extend the stochastic process models to achieve higher prediction accuracy, e.g., relax the lack of memory assumption, add nonstationarity, etc.

REFERENCES

Abe, M. (2009). "counting your customers" one by one: A hierarchical bayes extension to the pareto/nbd model. *Marketing Science*, 28(3):541–553.

Bakshy, E., Eckles, D., and Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pages 283–292. ACM.

Dahana, W. D., Miwa, Y., and Morisada, M. (2019). Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market. *Journal of Business Research*, 99:319–331.

Dechant, A., Spann, M., and Becker, J. U. (2019). Positive customer churn: An application to online dating. *Journal of Service Research*, 22(1):90–100.

Dmitriev, P., Frasca, B., Gupta, S., Kohavi, R., and Vaz, G. (2016). Pitfalls of long-term online controlled experiments. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1367–1376. IEEE.

Dziurzynski, L., Wadsworth, E., and McCarthy, D. (2014). BTYD: Implementing Buy 'Til You Die Models. *URL http://CRAN.R-project.org/package=BTYD. R package version*, 2.

Ehrenberg, A. S. C. (1972). *Repeat-buying; theory and applications*. North-Holland Pub. Co.

Fader, P. S. and Hardie, B. G. (2007). How to project customer retention. *Journal of Interactive Marketing*, 21(1):76–90.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2005). "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing science*, 24(2):275–284.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2008). Computing p(alive) using the bg/nbd model.

Fader, P. S., Hardie, B. G., and Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, 29(6):1086–1108.

Gomez-Uribe, C. A. and Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13.

Hohnhold, H., O'Brien, D., and Tang, D. (2015). Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858. ACM.

Karlin, S. (2014). *A first course in stochastic processes*. Academic press.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM.

Kohavi, R., Henne, R. M., and Sommerfield, D. (2007). Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967. ACM.

Lee, M. R. and Shen, M. (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 491–499. ACM.

Roy, R. K. (2001). *Design of experiments using the Taguchi approach: 16 steps to product and process improvement*. John Wiley & Sons.

Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who-are they and what will they do next? *Management science*, 33(1):1–24.

Tang, D., Agarwal, A., O'Brien, D., and Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26. ACM.

Thomke, S. H. (2003). *Experimentation matters: unlocking the potential of new technologies for innovation*. Harvard Business Press.

Trinh, G. (2013). *Modelling changes in buyer purchasing behaviour*. PhD thesis, University of South Australia.

Venkatesan, R., Bleier, A., Reinartz, W., and Ravishanker, N. (2019). Improving customer profit predictions with customer mindset metrics through multiple overimputation. *Journal of the Academy of Marketing Science*, 47(5):771–794.