Applying Shifted-Beta-Geometric and Beta-Discrete-Weibull Models for Employee Retention Curve Projection

Evgeny A. Antipov

eugene.antipov@gmail.com

Canadian University Dubai

Anastasia Gagarskaya

National Research University Higher School of Economics

Yulia Trofimova

National Research University Higher School of Economics

Elena Pokryshevskaya

Dubai Internet City

Research Article

Keywords: employee churn, churn prediction, probabilistic models, beta distribution, geometric distribution, Weibull distribution

Posted Date: August 16th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4765185/v1

Additional Declarations: No competing interests reported.

Abstract

Employees are vital assets to any organization, and their departure can result in reduced human capital and operational disruptions. To mitigate this, companies employ predictive analysis to forecast potential employee churn. Probability-based modeling for projecting employee churn is an underexplored area in HR analytics. This paper tests the applicability of the shifted-beta-geometric (sBG) and beta-discrete-Weibull (BdW) models within the context of employee survival projection. Using data from three cohorts of employees, we compare the results of these models with each other as well as with linear and logarithmic regressions. Our key finding is the superior performance of the BdW model, which can capture differences in churn rates between employees and within employees over time. The beta distribution captures the heterogeneous employee loyalty, while the Weibull distribution effectively captures retention rate changes over time. Our research demonstrates that parsimonious probabilistic models, which require minimal data and have so far been used only in customer analytics, can be applied in HR analytics for projecting employee retention curves.

1. Introduction

Churn prediction is a significant focus of customer analytics research (Yiğit & Shourabizadeh, 2017), but employee churn, which incurs substantial losses for companies (Saridakis & Cooper, 2016), receives less attention (Saradhi & Palshikar, 2011; Yiğit & Shourabizadeh, 2017). Predicting employee churn involves evaluating turnover rates over specific periods (Gentek, 2022). Employee churn represents a loss of intellectual assets for a company (Musanga & Chibaya, 2023), leading to operational disruptions, knowledge loss, and increased hiring costs (Winne et al., 2018). High turnover requires constant recruitment, but the irreplaceability of experienced employees exacerbates the issue (Suraihi et al., 2021). The significant expenses of adaptation, training, and salaries contribute to these losses (Naz et al., 2022). Thus, organizations seek to predict turnover and enhance retention strategies (Musanga & Chibaya, 2023). Accurate predictions enable companies to improve retention strategies and maintain a stable workforce. Predictive models for employee churn can help prevent workforce loss and provide a competitive advantage. At the same time, many businesses are unfamiliar with effective churn projection approaches (Musanga & Chibaya, 2023).

Job quitting and customer churn in contractual settings are similar stochastic processes. Both involve individuals leaving their group at the end of a period (Alamsyah & Salma, 2018; Gentek, 2022). Key metrics in these processes include retention and churn rates. The retention rate measures the proportion of individuals who remain active throughout the period, while the churn rate measures the proportion who leave by the end (Fader & Hardie, 2007).

Existing literature categorizes churn prediction models into two broad groups: machine learning approaches and probabilistic models based on statistical principles (Gentek, 2022). Various studies compare these models, with machine learning and predictive analytics being commonly used for their ability to process large datasets and capture complex relationships (Fader & Hardie, 2018).

Commonly used machine learning models include traditional classification algorithms (e.g., Logistic Regression, Random Forest, Naive Bayes, SVM, Decision Trees) (Kim et al., 2008; Alamsyah & Salma, 2018; Singh et al., 2019; Nestor et al., 2019; Gentek, 2022), deep learning models (Mena et al., 2019; Ozcan & Ozmen, 2021), and predictive analysis with time series models (e.g., ARIMA, ETS, XGB, LightGBM) (Jung, 2011; Javed & Azhar, 2017; Gregory, 2018). These models require large, representative datasets and often struggle with non-linear relationships and changing behaviors (Naz et al., 2022; Nestor et al., 2019; Kang & Oh, 2023).

In many cases, limited data makes traditional predictive models impractical. Therefore, this research focuses on probabilistic models for churn or retention projection that do not require extensive data but can be applied to short time series of employee cohort sizes over time. Probabilistic models, such as time series analysis, Bayesian models, survival analysis, logistic regression, SDEs, Poisson models, and exponential distributions, use probability distributions to project individual behaviors into the future (Tamaddoni et al., 2016).

Probabilistic models offer advantages over machine learning approaches, including ease of implementation, transparency, and stability with limited data (Chakraborty et al., 2015). Most studies on probabilistic models focus on customers, medical science, finance, demographics, politics, weather, and sports (Gandy, 2012; Berry et al., 2020; Zhang & Thomas, 2012; Jun et al., 2016; Kolasa & Rubaszek, 2012; Zhou et al., 2022; Alho, 2014; Levene & Fenner, 2021; Iversen et al., 2016; Boshnakov et al., 2017). Fader and Hardie's probabilistic models, primarily used for customer churn, retention, or lifetime value prediction, provide a theoretical framework based on probabilistic and statistical principles (Fader & Hardie, 2018). These models offer valuable insights for understanding customer behavior and can be adapted for employee churn/retention forecasting. This study contributes to the literature by exploring the application of Fader and Hardie's shifted-beta geometric (sBG) and Beta-discrete Weibull (BdW) models (originally developed for customer analytics) to the case of employee churn prediction. We compare the out-of-sample performance of these probabilistic models using a logarithmic trend model as a benchmark. The results offer practical insights for HR analytics, presenting an alternative solution for employee churn prediction that is easy to implement without machine learning.

The remainder of this study is organized as follows. The next section describes the chosen datasets. Section 3 presents the models applied in the study and compares them using multiple training/testing splits. Sections 4 and 5 discuss the results, findings, and conclusions, providing academic and managerial insights and proposing the most suitable model for specific cases.

2. Data

Our research leverages two unique US open data sources, typically not publicly available, offering an unparalleled opportunity to study the dynamics of employee cohort size. The first dataset provides information on regular hire employees within the County of Marin government organization, covering various departments from 2012 to 2020. The second dataset includes Baton Rouge's City-Parish employees from 2005 to 2017, encompassing employment status until transitioning to a new payroll system. This dataset spans multiple departments, enabling a more comprehensive analysis.

We examined the dynamics of three cohorts by tracking the number of employees remaining in each cohort over time:

1. **Dataset 1**: This dataset contains the monthly dynamics of a cohort from the County of Marin's Public Works Department, registered to be hired in December 2012 (initial size: 1275 employees) from December 2012 to December 2018. A noteworthy feature of this dataset is that the December 2012 cohort includes not only those hired in December 2012 itself but also those hired in previous months who were still active in December 2012. While this cohort definition is non-standard, it is practical for employers who rarely have large monthly cohorts but wish to project the monthly dynamics of employees active as of a specific date.
2. **Dataset 2**: This dataset details the monthly dynamics of a cohort from the Baton Rouge Police Department, hired in December 2005 (initial size: 53 employees) from December 2005 to December 2017. This cohort uses a standard definition, consisting of individuals hired within the same month.
3. **Dataset 3**: Similar to Dataset 2, this dataset includes the monthly dynamics of a cohort from the Baton Rouge Public Works Department, hired in December 2005 (initial size: 45 employees) from December 2005 to December 2017, also using a standard cohort definition.

By analyzing these datasets, we aim to gain insights into the retention and turnover trends within different government departments, thereby contributing valuable knowledge to workforce management and planning.

3. Methods

3.1 sBG Model by Fader & Hardie (2007)

The shifted-beta-geometric (sBG) distribution, introduced by Fader & Hardie (2007), is a model designed to predict customer retention rates based on single cohort data. This model operates within a hypothetical contractual framework, examining annual retention rates for cohorts of individuals. As a discrete-time model for contract duration, the sBG forecasts future churn rates within a cohort by analyzing past churn data (Fader & Hardie, 2009). Unlike other probability models such as Bayesian approaches or survival analysis methods like Kaplan-Meier estimation or Cox proportional hazards model, the sBG model requires limited data and involves only a few calculations, making it implementable with standard Microsoft Excel functions.

The sBG model aims to estimate the alpha and beta parameters of the beta distribution that maximize the log-likelihood, thus measuring the estimated churn rates. It integrates Beta and Geometric distributions. The Beta distribution accounts for individual heterogeneity (e.g., differences in retention rates and propensities to stay or leave the company) and shifts with each change in time, reflecting the dynamic nature of employee behavior over time. The Geometric distribution models the probability of churning after a period, projecting the number of periods it takes for the event (e.g., employee termination) to occur (Ahn et al., 2020).

As the authors of the model state (Fader & Hardie, 2007), the sBG model is based on two main concepts. The first one is expressed by the Beta-distribution, by which heterogeneity in churn probability *θ* is presented as:

$$f\left(\theta \,|\alpha\,,\,\beta\,\right)=\frac{\theta^{\,\alpha\,-1}(1-\theta\,)^{\beta\,-1}}{B(\alpha\,,\,\beta\,)},\;\;\alpha\,,\,\beta\;>0,$$

where:

α and *β* stand for the parameters that shape the beta distribution and capture the heterogeneity of observed individuals' behavior;

θ represents the probability of an event to happen (e.g., customer churns, employee leaves a company);

f(θ|α, β) represents the probability density function of beta distribution for the θ parameter with shape parameters alpha and beta;

B(α, β) represents the normalization constant that guarantees that the range under the probability density curve sums to 1.

The second one involves the Geometric distribution and describes that an employee does not churn and stays in the company with continuous retention probability *1 – θ*. The duration of relationships between an employee and a company is characterized by the shifted geometric distribution that can be expressed as the following:

$$P\left(T=t|\theta\,\right)=\theta\left(1-\theta\,\right)^{t-1},\;\;t=1,\;2,\;3,\;\ldots$$

$$S=\left(t|\theta\,\right)=\left(1-\theta\,\right)^{t},\qquad t=1,\;2,\;3,\;\ldots$$

,

where:

θ stands for the probability of an individual to stay active;

t stands for the number of trials until an event happens;

represents the probability of an individual to churn after each trial;

1-θ represents the probability of an individual to churn after each trial;
P(T = t / θ) stands for the probability of an event to happen (e.g., an individual to churn) after *t* of successful trials (e.g., months) given a retention probability of θ.

The model also considers that it is impossible to use the formulas mentioned above directly since the value of *θ* is unknown. *θ* parameter represents a value, which stands for the probability of an employee dropping out during a certain period. The parameter is used in the shifted-geometric distribution to model the likelihood of churn. To solve the issue (*θ* — unknown), Fader and Hardie used the mathematical expectation of formulas for the beta distribution, which characterizes the heterogeneity of the cross-section. In other words, since the *θ* parameter is assumed to vary across the population, the overall probability is computed by considering the expected value based on a beta distribution for *θ*. This allows the model to display the corresponding result for a randomly selected person. The probability mass function and survivor function of the sBG model can be expressed as the following (Fader & Hardie, 2007):

$$P\,(T = t|\alpha\ ,\ \beta\ ) = \frac{B(\alpha\ + 1,\ \beta\ + t - 1)}{B(\alpha\ ,\ \beta\ )},\ t = 1,\ 2,\ \ldots$$

$$S\,(t|\alpha\ ,\ \beta\ ) = \frac{B(\alpha\ ,\ \beta\ + t)}{B(\alpha\ ,\ \beta\ )},\ \ t = 1,\ 2,\ \ldots\ ,$$

where:

P(T = t/ α, *β*) represents the probability of a successful event to happen (e.g., customer makes a purchase);

S(*t/ α*, *β*) represents the probability of a successful event to occur given that it has not occurred yet (e.g., we are predicting that a customer will make their 1, 2, … purchase during the next transaction, given that they have not made it yet);

α and *β* are the parameters that shape the beta distribution and capture observed individuals' behavior heterogeneity.

Additionally, the authors mention that sBG probabilities can be measured by application of the following formula without the direct involvement of the beta functions through a forward-recursion formula:

$$P\,(T = t|\alpha\ ,\ \beta\ ) = \begin{cases} \frac{\alpha}{\alpha\ +\beta} & t = 1 \\ \frac{\beta\ +t-2}{\alpha\ +\beta\ +t-1} P\,(T = t - 1) & t = 2,\ 3,\ \ldots \end{cases}$$

Since working with beta functions might be complicated, the forward-recursion formula simplifies the model implementation process. It allows us to calculate the churn probability at the beginning *(T = 1)* and then iteratively calculate probabilities for further periods. If the *P(T = t)* is known, it is possible to use the following formula for the survivor function calculation, which describes the probability of an individual's survival:

$$S\,(t|\alpha\ ,\ \beta\ ) = 1 - \sum\ {}_{i-1}^{t} P(T = i|\alpha\ ,\ \beta\ ),$$

where:

$\sum{}_{i-1}^{t} P(T = i|\alpha\ ,\ \beta\ )$ represents the sum of cumulative probabilities of successful events to occur once, twice, thrice, and up to t-times.

3.2 BdW Model by Fader & Hardie (2018)

In 2018, Fader and Hardie introduced the beta-discrete-Weibull (BdW) model, an extension of the beta-geometric (BG) model. The primary distinction between the two models lies in their handling of individual-level churn probability fluctuations (Fader & Hardie, 2018). The sBG model addresses increasing cohort retention rates over time by accounting for cross-sectional heterogeneity (Fader & Hardie, 2007). However, in a continuous time setting, retention rate changes can be better explained by varying tendencies to churn among individuals rather than by increasing periods alone. This individual diversity reflects changes in retention indicators, and the BdW model captures these individual-level retention probabilities, which may increase or decrease over time (Vijayaragunathan & Kishore, 2022).

Fader and Hardie emphasized that while cohort-based ratios generally tend to increase monotonically over time, there are exceptions. For example, some cohorts experience an initial drop before the retention rate increases. The BG distribution cannot adequately model such patterns, prompting the development of the BdW model. The BdW model's flexibility allows it to accommodate non-monotonically increasing retention rates by accounting for individual-level changes in churn tendencies (Fader & Hardie, 2018). The Weibull distribution can model both increasing and decreasing churn probabilities over time for individuals within a cohort, thus providing a more nuanced understanding of retention dynamics.

The BdW model's conceptual foundation lies in the discrete-Weibull (dW) distribution. Fader and Hardie (2018) reviewed existing theories and research on the dW distribution, noting that later dW models (Murthy et al., 2004; Rinne, 2009) offer simplicity and flexibility for discrete-time analysis. The BdW model combines aspects of both geometric and Weibull distributions, making it duration-dependent.

However, in testing with "Regular" and "High End" datasets of contractual settings (Berry and Linoff, 2004), Fader and Hardie (2018) found that the dW distribution alone could not accurately forecast the number of surviving individuals within a cohort. Conversely, the BG model, which accounts for heterogeneity, better predicted retention rates. Their findings suggest that an effective churn prediction model must capture both the length of the period and individual heterogeneity for improved accuracy. Consequently, the beta function, integral to the beta distribution, was incorporated to enhance the model's predictive capability.

In summary, the BdW model extends the BG model by integrating the flexibility of the Weibull distribution to handle diverse retention patterns within cohorts. This approach provides a more accurate tool for predicting churn, particularly in settings where individual-level retention probabilities vary significantly over time.

The following equation is described as parametric mixture model the beta-discrete-Weibull:

$$P\,(T = t|\gamma\ ,\ \delta\ ,\ c) = S\,(t-1|\gamma\ ,\ \delta\ ,\ c) - S\,(t|\gamma\ ,\ \delta\ ,\ c)$$

$$= \frac{B\left(\gamma\,,\,\delta\,+(t-1)^{c}\right)-B\left(\gamma\,,\,\delta\,+\,t^{c}\right)}{B\left(\gamma\,,\,\delta\,\right)},\,t=1,\,2,\,3,\,\ldots\,,$$

where:

S(t-1 | γ, δ, c) and *S(t | γ, δ, c)* represent survival functions at times (t-1) and (t), respectively;

B(γ, δ + (t-1) ^{c} *)* and B(*γ, δ + (t)*^{c}*)* represent cumulative distribution functions (CDF) of beta-distribution, including gamma and delta as parameters (CDF represents the probability to churn by time *t-1* and *t*);

B(γ, δ) represents the normalization factor of the overall cumulative distribution function (probability to churn) of the beta distribution.

3.3 Linear and log-linear regression models

In addition to the sBG and BdW models, linear and semi-logarithmic (from now on, “logarithmic” for brevity) analysis linking the retained cohort % (y) and the period (t) is applied in this research as simple benchmarks. The following two specifications were used:

$$y=a+bt\;\text{and}\;y=a+bln\left(t\right)$$

3.5 Model comparison

To evaluate the effectiveness of the models, we predicted employee retention for periods of 1 year, 2 years, and 4 years across the three datasets. The forecasted variable was the percentage share of the cohort retained in a given month. The testing sample’s mean absolute error (measured in percentage points) was used to compare models.

Each model was tested using multiple training/testing splits:

- For dataset 1: 61 months / 12 months, 49 months / 24 months, 25 months / 48 months.
- For dataset 2 and dataset 3: 133 months / 12 months, 121 months / 24 months, 97 months / 48 months.

These splits allow us to assess the models' performance across different time horizons and ensure a robust comparison of their predictive capabilities.

4. Results

We aimed to determine the most universally effective model for predicting employee survival across different scenarios, considering that the dynamics of employee numbers can vary. To achieve this, we collected and analyzed data from three distinct datasets, each representing employees within different organizational contexts. Models were evaluated based on their ability to predict employee survival over varying time horizons, using different lengths of training and testing periods (Table 1).

Table 1
Testing sample's MAE for the sBG, BdW, Linear, and Logarithmic models for different training/testing splits.

	The number of months used for training the model	The number of months used for testing the model	sBG model's MAE, %	Linear model's MAE, %	BdW model's MAE, %	Logarithmic model's MAE, %
Dataset 1	61	12	5.22	3.47	1.54	8.31
	49	24	3.58	2.24	0.40	8.36
	25	48	6.02	5.20	2.14	8.28
Dataset 2	133	12	10.80	7.51	1.85	1.59
	121	24	12.53	10.32	1.48	1.59
	97	48	14.59	19.61	1.86	3.84
Dataset 3	133	12	4.90	3.42	1.91	2.37
	121	24	5.18	3.59	2.61	2.78
	97	48	5.20	4.94	5.19	5.04

The BdW model systematically performed much better than other models except for dataset 3, where the logarithmic model performed similarly. During the longest testing period, the MAE values for all models were approximately the same in the case of dataset 3.

Overall, the BdW model is the most versatile and best-performing model for all datasets and forms of dynamics, exhibiting fewer errors than other models. The advantage of the best-performing model is most pronounced over longer forecasting horizons. This is particularly evident in the second dataset, where the BdW model consistently outperforms the other models, especially when predicting retention over 48 months. This suggests that the BdW model offers a distinct advantage in capturing and accurately predicting underlying trends and patterns within the data for longer-term forecasting needs, such as predicting employee retention over several years.

Next, to assess the reliability of the BdW model in predicting employee churn and determine the acceptable number of months required to train the model, we fixed the number of months used for testing and incrementally increased the training period by 12 months. We then calculated the MAE for predictions made 12, 24, 36, and 48 months forward. The obtained results are presented in Table 2, and plots of the dependence of MAE on training sample size for each test sample size and for each of the three datasets are shown in Fig. 1.

Table 2
The BdW model estimation results for the longest training/testing splits.

The number of months used for training the model	The number of months used for testing the model	BdW model's MAE, %		
		Dataset 1	Dataset 2	Dataset 3
12	12	0.66	5.61	1.88
24	12	0.25	7.23	1.26
36	12	0.24	5.33	0.65
48	12	0.81	3.55	0.78
60	12	0.54	3.64	0.84
12	24	1.37	4.62	1.50
24	24	1.22	13.37	0.96
36	24	0.32	7.93	1.48
48	24	0.74	6.28	1.19
60	24	–	3.79	0.55
12	36	2.40	8.07	1.20
24	36	1.83	18.45	1.06
36	36	0.65	11.13	2.04
48	36	–	7.60	1.13
60	36	–	4.57	0.62
12	48	3.71	12.21	1.46
24	48	2.14	23.33	1.32
36	48	–	13.32	2.19
48	48	–	8.99	1.34
60	48	–	5.86	1.20

The comparison results generally imply a tendency for larger training sets to ensure lower testing errors. For datasets 1 and 2, the most significant drop in MAE is observed when the training sample increases from 24 to 36 months. The MAE's dependence on the training sample size is least pronounced in the case of dataset 3, which is characterized by the lowest and most homogeneous MAEs across all experiments. Our comparison shows that even using as few as 12 months of training data provides decent accuracy for 12-month ahead forecasts, confirming the ability of the chosen theoretically justified BdW model to extrapolate many periods ahead.

5. Discussion

The BdW model has shown the most uniformly good performance across several datasets and is considered a reliable tool for long-term retention projections. Employee retention projection is a complex process that involves both the duration and estimation of behavioral factors (e.g., reasons for churn, robust external factors at certain times) to achieve accurate outcomes. Therefore, the model must consider increasing periods and capture the sample's heterogeneity.

According to Fader and Hardie (2018), retention rates change over time due to the persistently changing behavior of customers, which similarly applies to employees' behavior and reasons for churn. The BdW model performed the best among the tested models because it can capture individual-level churn/retention probabilities over increasing time durations.

The poorer performance of other models can be attributed to their inability to grasp the relationships between variables affecting potential retention rate changes over time. For instance, the sBG model presented by Fader and Hardie (2007) showed satisfactory results for short-term projections but may not perform robustly with other datasets over longer periods. The sBG model includes Beta and Geometric distributions, considering both sample heterogeneity and the probability of churning over time. However, it lacks the flexibility to capture long-term behavior tendencies effectively. In contrast, the BdW model offers a more flexible distribution projection for event occurrence. Unlike geometric distribution, the Weibull distribution captures nuances in individual behavior patterns over time (e.g., the likelihood of an employee staying or leaving a company).

The linear and logarithmic regression models also fall short of long-term predictions. The linear model assumes a linear trend in cohort size, which does not hold for the complex dynamics of employee retention. Retention projection involves capturing non-linear relationships and evaluating evolving factors (e.g., external environment changes, HR market dynamics, and employees' physical and mental states). While logarithmic regression can capture nonlinear changes, it poorly corresponds to the underlying data-generating process.

Fader and Hardie (2018) highlight the BdW model's flexibility in handling inhomogeneous changes in retention rate projection. Several studies emphasize the Weibull distribution's flexibility in capturing complex survival patterns and suggest different extensions to develop more accurate predictive models. For example, Jamal and Bucklin (2006) found that the Weibull model outperformed simpler hazard modeling approaches, underscoring the importance of integrating heterogeneity for churn/retention projection. Enkhmunkh et al. (2007) highlighted the Weibull distribution's accuracy and flexibility for shape-forming and simple cumulative distribution functions. Nekoukhou et al. (2017) introduced the discrete equivalent of the beta-Weibull distribution for discrete data and survival modeling. Chanasriphum et al. (2019) developed a model based on the beta-generalized Weibull distribution for lifetime data forecasting, which outperformed other regression models.

6. Conclusion

This paper investigates the feasibility of applying the BdW model, originally developed by Fader and Hardie (2018) for customer analytics, to project employee retention curves. We compared the accuracy of two probabilistic models—the sBG and BdW models—originally designed for customer churn prediction, to assess their suitability for employee churn prediction. Testing these models on three real-life datasets, we found that both models are effective in HR analytics for predicting employee retention, with the BdW model outperforming the sBG, the linear, and logarithmic regression models by capturing changes in attrition rate heterogeneity over time. Our analysis indicates that using 48 months of historical data can provide accurate retention rate predictions for future periods.

Using probabilistic models for employee retention curve projection is addressed for the first time in academic HR analytics literature. Our research adds scientific novelty by demonstrating that such parsimonious models based on limited data can give surprisingly accurate forecasts. Although ML approaches can handle large datasets and generate individual-level predictions, they often lack interpretability, require substantial data, and apply mostly to short-term forecasting (e.g., 1–6 months ahead). On the other hand, probabilistic models offer greater transparency of underlying data-generating processes. They can be implemented using tools like Microsoft Excel, making them accessible for organizations with limited data science expertise. Firms can make projections for many months ahead and plan recruitment accordingly. While model-based estimates naturally cannot account for rare external shocks, they can serve as a useful diagnostic tool for detecting the company's under- or over-performance from the employee retention perspective.

However, probabilistic models have their limitations. Unlike customer churn, employee turnover can be voluntary or involuntary, varying by circumstances and industries. These different churn types may require distinct approaches, affecting forecast quality. Additionally, the presented parsimonious probabilistic models apply mostly to cohorts of relatively homogeneous employees (regarding the job function).

Future research can test and evaluate the accuracy of these results across different industries and role types. Given the rapidly changing environments of recent years, it would be useful to test the models with recent datasets to determine if high degrees of uncertainty affect the quality of retention projections or if the best probabilistic models still generalize well even without explanatory variables.

Declarations

Competing Interests

None

Funding

None

Ethics approval/declarations (include appropriate approvals or waivers)

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The spreadsheet with data and model implementations is available in the dedicated Open Science Framework (OSF) repository.

Code availability (software application or custom code)

Not applicable

Authors' contributions

E.A. conceptualized, planned, and reviewed the manuscript. A.G. developed the spreadsheet models. Y.T. prepared the data and wrote the main text. E.P. wrote the conclusion and extensively reviewed the whole paper.

References

1. Ahn J, Hwang J, Kim D, Choi H, Kang S (2020) A survey on churn analysis in various business domains. IEEE Access 8:220816–220839
2. Alamsyah A, Salma N (2018) A comparative study of employee churn Prediction model. 2018 4th International Conference on Science and Technology (ICST)
3. Alho JM (2014) Forecasting demographic forecasts. Int J Forecast 30(4):1128–1135
4. Al-Suraihi WA, Samikon SA, Alsuraihi A, Ibrahim I (2021) Employee Turnover: Causes, importance and retention strategies. Eur J Bus Manage Res 6(3):1–10
5. Berry LR, Helman P, West M (2020) Probabilistic forecasting of heterogeneous consumer transaction–sales time series. Int J Forecast 36(2):552–569
6. Boshnakov GN, Kharrat T, McHale IG (2017) A bivariate Weibull count model for forecasting association football scores. Int J Forecast 33(2):458–466
7. Chakraborty M, Das S, Lavoie A (2015) How to show a probabilistic model is better. arXiv (Cornell University)
8. Chanasriphum N, Seenoi P, Srisodaphol W (2019) The Log Beta generalized Weibull regression model for lifetime data. Journal of Physics: Conference Series, 1366(1), 012121
9. De Winne S, Marescaux E, Sels L, Van Beveren I, Vanormelingen S (2018) The impact of employee turnover and turnover volatility on labor productivity: a flexible non-linear approach. Int J Hum Resource Manage 30(21):3049–3079

10. Enkhmunkh N, Kim G, Hwang K, Hyun S (2007) A parameter estimation of Weibull distribution for reliability assessment with limited failure data. International Forum on Strategic Technology, 39–42. Ulaanbaatar, Mongolia
11. Fader PS, Hardie BGS (2009) Probability Models for Customer-Base Analysis. J Interact Mark 23(1):61–69
12. Fader PS, Hardie BGS, Lee KL (2005) Counting your customers the easy way: an alternative to the Pareto/NBD model. Mark Sci 24(2):275–284
13. Fader PS, Hardie BGS Fitting the sBG Model to Multi-Cohort Data., BruceHardie (2007) https://www.brucehardie.com/notes/017/. Accessed on 15 December 2023
14. Fader PS, Hardie BGS (2007) How to Project Customer Retention. J Interact Mark 21:76–90
15. Gandy A (2012) Performance monitoring of credit portfolios using survival analysis. Int J Forecast 28(1):139–144
16. Gentek A (2022) Employee Churn Prediction in Healthcare Industry using Supervised Machine Learning (MA thesis). KTH Royal Institute of Technology. Retrieved from https://www.diva-portal.org/smash/get/diva2:1711505/FULLTEXT01.pdf
17. Gregory B (2018) Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data. arXiv (Cornell University
18. Iversen EB, Morales JM, Møller JK, Madsen H (2016) Short-term probabilistic forecasting of wind speed using stochastic differential equations. Int J Forecast 32(3):981–990
19. Jamal Z, Bucklin RE (1987) Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. J Interact Mark 20(3–4):16–29
20. Javed S, Azhar A (2017) Forecasting Employee Turnover for Human Resource Based on Time Series Analysis. Int J Econ Res 14(16):445–456
21. Jun DB, Kim K, Park M (2016) Forecasting annual lung and bronchus cancer deaths using individual survival times. Int J Forecast 32(1):168–179
22. Kang S, Oh H (2024) Forecasting South Korea’s presidential election via multiparty dynamic Bayesian modeling. Int J Forecast 40(1):124–141
23. Kim D, Lee H, Cho S (2008) Response modeling with support vector regression. Expert Syst Appl 34(2):1102–1108
24. Kolasa M, Rubaszek M (2015) Forecasting using DSGE models with financial frictions. Int J Forecast 31(1):1–19
25. Levene M, Fenner T (2021) A stochastic differential equation approach to the analysis of the 2017 and 2019 UK general election polls. Int J Forecast 37(3):1227–1234
26. Mahsa EG, Jafar TM (2013) Customer Lifetime Value Models: A literature Survey. Int J Industrial Eng Prod Res 24(4):317–336
27. Masarifoglu M, Büyüklü AH (2019) Applying survival analysis to telecom churn data. Am J Theoretical Appl Stat 8(6):261

28. Mena CG, De Caigny A, Coussement K, De Bock KW, Lessmann S (2019) Churn Prediction with Sequential Data and Deep Neural Networks. A Comparative Analysis. arXiv (Cornell University)
29. Naz K, Siddiqui IF, Koo J, Khan MA, Qureshi NMF (2022) Predictive modeling of employee churn analysis for IoT-Enabled software industry. Appl Sci 12(20):10495
30. Nekoukhou V, Bidram H, Roozegar R (2016) The Beta-Weibull Distribution on the Lattice of Integers. Ciência E Natura 39(1):40
31. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, Goldenberg A, Ghassemi M (2019) Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. Proceedings of Machine Learning Research, 106, 1–23
32. Saradhi VV, Palshikar GK (2011) Employee churn prediction. Expert Syst Appl 38(3):1999–2006
33. Saridakis G, Cooper CL (2016) Introduction: the state of employee turnover. In Edw Elgar Publishing eBooks, 1–4
34. Singh P, Singh SP, Singh DS (2019) An Introduction and Review on Machine Learning Application in Medicine and Healthcare. IEEE Conference on Information and Communication Technology, 1–6
35. Tamaddoni A, Stakhovych S, Ewing MT (2015) Comparing churn prediction techniques and assessing their performance. J Service Res 19(2):123–141
36. Yigit IO, Shourabizadeh H (2017) An approach for predicting employee churn by using data mining. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). Malatya, Turkey

Footnotes

[1] Employees Jobs and Attributes County of Marin

[2] Legacy City-Parish Employees

[3] % in the context of this paper indicates percentage points. For example, if 80% of the cohort were retained in period *t*, while the model prediction is 78%, the MAE is two percentage points.

Figures

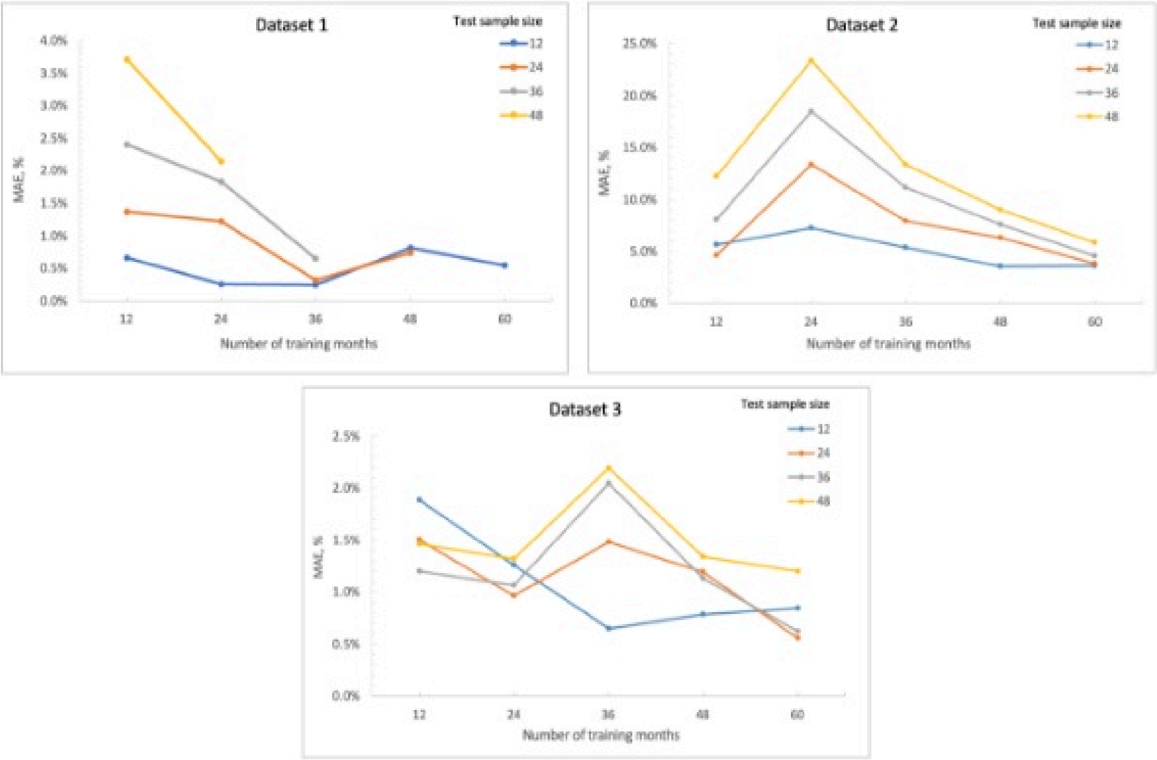

Figure 1

Dependence of MAE (%) on training sample size for each test sample size and for each of the three datasets