Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38785

Using Regression and Machine Learning to Forecast Customer Spending Patterns and Turnover

Hao Zhang

Palo Alto High School, Palo Alto, USA
hz55340@pausd.us

Abstract. The study is based on a customer database with 440,832 valid data and uses the following methods: Matrix based Regression and Machine Learning. The first goal is to see if there is a correlation between the customer's age and the number of dollars spent, and if there is a correlation, is there any gender trend? The mathematical underpinning is introduced via the least squares residual minimization, matrix conditioning and the full-column-rank condition for uniqueness. Using the entire cleaned dataset, different linear and quadratic models are fit for female customers and male customers. Age trends are small and negative in both linear models (R-squared values less than 0.01); quadratic models do not help much. A separate computational-cost exercise checks the first 1, 000 observations and finds a balanced gender composition of 51.4% female and 48.6% male. The second will involve prediction of turnover based on age, gender, duration of engagement, purchases, number of interactions for support, and amount spent. The ROC-AUC for logistic regression is 0.906 and the ROC-AUC for random forest is 0.967. The best practical indicator is the turnover rate which rises sharply when support contacts become more numerous. Results show that the age spending regression has weak explanatory structure, whereas turnover is strongly modeled in a multivariate framework.

Keywords: Least squares, Customer churn, Logistic regression, Random forest, Customer support

1. Introduction

Many organizations are gathering huge amounts of customer data, including demographic data, purchase history, customer engagement metrics and service history. These data open up the possibility to shift the Analytics paradigm from descriptive reporting to quantitative models of customer behavior. There are two questions in this data set that are particularly effective. First, is there any simple demographic variable, like age, which can explain how much the customer is spending and is there any difference between females and males? Secondly, is it possible to use the same database to find out which customers are likely to terminate their association with the company? The first question is obviously a least-squares regression problem and the second is binary prediction problem. They demonstrate the use of matrix computation to explain and predict, in a realistic business environment.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38785

Least squares remain one of the central tools of numerical linear algebra because it expresses data fitting through the minimization of a residual norm. Recent numerical work continues to emphasize that the mathematical formulation and the computational method should not be confused. Wathen notes that the normal equations are elegant but can create serious conditioning difficulties [1]. This is because forming A^TA squares the condition number of A. Meier et al. [2], who likewise study the numerical stability of modern large-scale least-squares solvers and show why stable solution strategies and preconditioning matter when problems become large. These concerns are relevant even when the number of predictors is small, because the project is intended to connect regression with matrix computation rather than to treat regression as a black-box statistical command.

Customer churn is another significant machine learning use case. Recent studies have revealed that churn is frequently considered as a binary classification problem and assessed by metrics like precision, recall, F1 score, and area under ROC curve. Prabadevi et al. make a comparison between logistic regression, random forest, K-nearest neighbors and gradient boosting for churn analysis [3]. Suh presents an extensive operational dataset from a rental-service company, and shows that information about customer behavior and interactions can be useful for prediction of churn [4]. A comparison of several machine-learning models and sampling strategies performed by Imani and Arabnia compared multiple machine-learning models and data-sampling strategies for customer churn prediction [5], while a 2024 telecommunications study identified random forest as its best-performing classifier [6]. Other recent studies have focused on time-varying customer behavior [7], on customer-support constraints in e-commerce [8], and on more sophisticated application of deep learning in the field of customer relationship management [9]. These studies inspire this project to compare the more flexible random forest with an interpretable logistic model.

Both spending analysis and turnover prediction are popular customer analytics applications, but they focus on different aspects of customer behavior. Spending regression is an example of a continuous outcome variable while turnover prediction is an example of a binary outcome variable. Comparing these two tasks in the same dataset can be used to determine if variables that have low explanatory power for one customer behavior can still be useful for the prediction of another. Specifically, age and gender are readily available demographic variables, but this does not mean that they are necessarily useful – and assuming they are in large datasets is a dangerous trap. Variables like purchase history, engagement duration, and support interactions may provide a more direct indication of a customer's relationship with a company.

This distinction provided the major research questions for the present study. The first question to ask is do the ages offer any useful explanatory power as a total customer spending model and do the models differ if gender-specific or nonlinear processes are examined. The second question is whether the combination of the two variables -- demographic and behavioral -- can result in significantly better customer turnover predictions. A third goal is to compare an interpretable statistical classifier, logistic regression, to a more flexible machine-learning technique, random forest. This study focuses not only on the presence or not of a fitted relationship, but also on the practical measures of usefulness, such as R-squared and RMSE for spending regression, precision, recall, F1 score and ROC-AUC for turnover prediction. By using this framework, the analysis can differentiate a mathematically detectable relationship from one that is strong enough to be used for practical prediction.

All 440,832 valid observations have been utilized in the main analysis of age-spending and turnover in this paper. The first 1, 000 observations are only utilized in the separate sampling question about computational cost. In Section 2, the theory of least-squares, conditioning problems, data preparation and modeling approach are presented. The linear and quadratic spending models,

DOI: 10.54254/2755-2721/2026.38785

the sampling check and the turnover models are reported in Section 3. Central finding is a contrast: Age explains very little of individual spending, while a signal of customer characteristics explains a lot of turnover risk. This contrast is important because a lack of usefulness in one regression does not suggest a lack of usefulness for the database as an entity; the usefulness of a variable depends on the thing being modeled.

The aim of this study is not methodological, but rather empirical and application-oriented. It does not suggest a new algorithm of regression or machine learning, but presents, using one large customer dataset, a practically significant contrast of results. Behavioral information in particular (esp. support interactions) supports much stronger turnover prediction than age and gender; age and gender have only weak explanatory power for individual spending. The analysis further formally tests gender differences in age-spending patterns compared to just looking at them or comparing separate models.

2. Method and theory

2.1. Least-squares regression and the residual norm

Let A be an m-by-n design matrix, x the vector of unknown coefficients, and b the observed response vector. A linear regression model seeks x such that Ax is close to b. The residual vector is $r = b - Ax$ [10, 11], and ordinary least squares selects the coefficient vector that minimizes the squared Euclidean, or 2-norm, residual of the form $\min_x \|b - Ax\|_2^2 = \Sigma_i(b_i - a_i^T x)^2$. The 2-norm is attractive for several reasons. Squaring the residuals produces a smooth differentiable objective, which makes the optimality conditions straightforward. It also gives a geometric interpretation: the fitted vector Ax -hat is the orthogonal projection of b onto the column space of A. Large residuals receive greater weight because they are squared, which is useful when large prediction errors should be penalized strongly. Under standard assumptions involving independent approximately Gaussian errors, least squares also have a familiar statistical interpretation. However, the same squaring makes the method sensitive to extreme observations. If a dataset contains severe outliers or heavy-tailed errors, minimizing the L1 norm, the sum of absolute residuals, can be more robust because a residual ten times larger receives ten rather than one hundred times the contribution.

2.2. The normal equations, conditioning and uniqueness

Differentiating the least-squares objective gives the normal equations $A^TA\ x = A^Tb$. If A has full column rank, A^TA is positive definite and invertible, and the least-squares coefficient vector is unique. Full column rank means that no predictor column can be expressed as a linear combination of the others. Equivalently, the null space of A contains only the zero vector. When A is rank deficient, fitted values can still be meaningful, but the coefficient vector need not be unique; a minimum-norm solution can instead be selected with the Moore-Penrose pseudoinverse [10, 11].

Conditioning describes how sensitive a numerical problem is to small perturbations. For the 2-norm, the condition number of a full-rank matrix is $\kappa_2(A) = \sigma \max(A)/\sigma\min(A)$, where the singular values measure stretching along orthogonal directions. A condition number near one indicates a well-conditioned problem, whereas a very large value means that small changes in the data or rounding errors can produce much larger changes in the solution [12]. The normal equations deserve special attention because $\kappa_2(A^TA) = \kappa_2(A)^2$ [1]. Thus, a design matrix with condition number 10^5 produces a normal matrix with condition number about 10^{10}. This is why QR

DOI: 10.54254/2755-2721/2026.38785

factorization or singular value decomposition is often preferred for numerically sensitive least-squares problems. Modern research on large least-squares systems continues to investigate stable alternatives and preconditioned methods [1, 2].

2.3. Dataset and spending models

The original customer file provided was 440,833 rows. The one row had no data at all, and was dropped, leaving 440,832 valid customer records. Those variables are customer ID, age, gender, total months that the customer has been with the company, number of purchases, number of support interactions, total dollars spent, and turnover status. Customer ID was not considered as a predictor as it is just used to identify the customer. The entire cleaned data set consists of 190, 580 female customers and 250, 252 male customers. The overall mean age is 39.37 years and the overall mean total spending is $631.62. More details can be found in Table 1.

Table 1. Summary statistics of cleaned customer data

Variable	Mean	Std. dev.	Min.	Median	Max.
Age	39.37	12.44	18	39	65
Months engaged	31.26	17.26	1	32	60
Purchases	15.81	8.59	1	16	30
Support interactions	3.60	3.07	0	3	10
Total spent ($)	631.62	240.80	100	661	1000

For the first applied question, age is the explanatory variable and total dollars spent is the response. Female and male customers are modeled separately so that both the intercept and age slope can differ across gender. The linear model is $y = \beta_0 + \beta_1 x$. A quadratic model, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, is then fitted to test whether a curved age-spending relationship provides a materially better description. R-squared measures the proportion of response variation explained by the model, while RMSE measures the typical prediction error in dollars. Because the dataset is extremely large, very small systematic patterns can be statistically detectable; therefore, practical effect size and model fit are more important here than merely observing a nonzero slope.

The separate descriptive fits are complemented by unified ordinary least-squares models to directly test for gender differences. The age variable is centered around the average age of the entire sample, which is 39.37 years. The linear interaction model also incorporates centered age, as well as an indicator for males and an age-by-gender interaction term; the interaction coefficient itself is a test of whether the age slope is different for the two sexes. The quadratic interaction model also includes centered age-squared and age-squared-by-gender. Confidence intervals for coefficients are given at 95% and the size of the practical effect is evaluated by the change in R-squared and not only by the statistical significance. Finally, an expanded spending model is used to investigate whether available customer characteristics explain spending better than age and gender does, using the quadratic interaction specification, but also including months engaged, purchases, and support interactions.

DOI: 10.54254/2755-2721/2026.38785

2.4. In the computational-cost question, the next few responses involve sampling over two rounds

The assignment also asks for an answer to: What might be an alternative size to the dataset if a manager desired a lower computational cost? The first 1,000 valid observations were examined as recommended by the professor before a sequential sample of larger size was examined. This check is not intended to serve as a substitute for the full data analysis, but rather to check if an initial reduced set of data is reasonably balanced with respect to the grouping variable in the spending comparison – gender. If the first 1, 000 observations were all one gender, then more observations would have been added and the ratio would be recalculated. For this simple screening procedure, the observed sample had 514 female and 486 male customers; the split was close enough. The sequential procedure follows the project guidance and a more formal production study could further investigate age, spend and turn-over distributions, or adopt random/stratified sampling.

So the 1, 000-record subset is merely a computational-cost screening exercise in the service of this project, to reduce by over 99% the number of records that need to be processed by this check, relative to the 440, 832-record cleaned dataset. Since its gender split is already quite close to balanced, there is no need to read further rows for that purpose only. In this paper, the complete cleaned data set is used to produce all inferential/predictive results.

2.5. The forecasting and assessment of turnover in companies

The turnover variable is a binary variable which is not best suited to ordinary linear regression because the regression can yield values outside the range of zero to one. Logistic regression rather models log-odds of turnover as a linear combination of predictors. Predictors considered here are age, gender, months engaged, purchases, support interactions and total spending. Numerical predictors are scaled with train-set statistics to prevent numerical scale dominance of numerical optimization. The data is split into training data (80%) and test data (20%) with a stratified split based on the turnover proportion. This separation is crucial as assessing on the training data would be an overestimate of the model's performance on unseen customers [13].

A random forest is used as a nonlinear comparison model. Random forests can be used to model threshold effects and interactions which are difficult to model in a linear or logistic model and which must be prescribed a priori. Recent churn literature highlights that logistic regression is commonly employed as an interpretable baseline model, and ensemble/tree-based models tend to exhibit greater predictive power. The performance is evaluated using accuracy, precision, recall, F1 score and ROC-AUC. When class frequencies vary, accuracy alone can be misleading, hence the importance of recall and precision for understanding missed churners and false alarms. ROC-AUC is a summary of the model's capacity to rank the churners above the non-churners at different thresholds of the classifications, and is used in many recent churn studies.

3. Results and applications

3.1. Linear age-spending results

As shown in Table 2, the female linear equation is $\hat{y} = 666.82 - 1.3423x$. Interpreted literally within the observed age range, one additional year of age is associated with about \$1.34 lower predicted total spending for female customers. The male equation is $\hat{y} = 716.50 - 1.8187x$, giving a somewhat steeper decrease of about \$1.82 per year. The intercepts should not be given a

DOI: 10.54254/2755-2721/2026.38785

direct business interpretation because age zero is outside the observed range of 18 to 65 years. The slopes in Figure 1 show a directional difference between genders, but the magnitude of model fit is much more important than the difference between the two slopes.

Table 2. Gender-specific linear and quadratic models of spending

Gender	Model	Equation	R^2	RMSE ($)	n
Female	Linear	$y = 666.82 - 1.3423x$	0.0048	246.80	190,580
Female	Quadratic	$y = 467.25 + 9.5453x - 0.13373x^2$	0.0129	245.80	190,580
Male	Linear	$y = 716.50 - 1.8187x$	0.0089	233.66	250,252
Male	Quadratic	$y = 446.97 + 12.9993x - 0.18482x^2$	0.0247	231.78	250,252

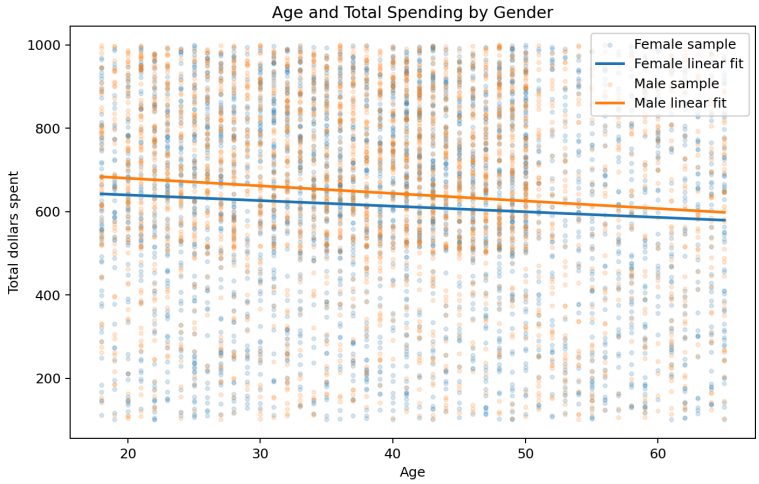


Figure 1. Total spending by gender and age. The regression lines represent the entire cleaned dataset; a subset of points is plotted for readability

R-squared for females is just 0.0048, so age accounts for about 0.48% of the variation in spending. The R squared value for males is 0.0089 (or about 0.89%). Both are less than one per cent. This is confirmed by the scatter plot, which shows that at almost every age there are customers who buy a wide range of products. Thus, the age of a customer is not a very relevant piece of information on the amount spent by the individual customer. The negative lines give an indication of weak overall trends, not of strong relationships. This is an important interpretation since a very large set of data can give the impression of a trend that is very smooth when the actual explanatory value of the trend may be small.

3.2. Quadratic models and practical fit

The quadratic models test if the age-spending relationship is curvilinear or not, if it is not linear. Both fitted x-squared terms are negative (refer to Figure 2), indicating that the curves have wide inverted-U shapes. For females, R-squared increases from 0.0048 to 0.0129 and RMSE decreases from $246.80 to $245.80. For males, R-squared increases from 0.0089 to 0.0247 and RMSE decreases from $233.66 to $231.78. The quadratic model thus somewhat better models the average pattern, particularly for males although the absolute fit is not very far improved. The most powerful age-only model accounts for just 2.47% of the total variance in spending.

DOI: 10.54254/2755-2721/2026.38785

As seen in this result, there is a distinction between model complexity and fruitful explanatory power. Including a polynomial term in the model will always add more flexibility, and can never make a poor in-sample least-squares fit worse; however, improving R-squared does not make the model useful. The RMSE here differs by just a dollar or two and the typical error is still more than $230. There is therefore some evidence for the slight bending, but not to the extent that it is considered a strong determinant of spending. One could use a higher order polynomial and get even closer to the observed averages, but this would also be more likely to model small fluctuations instead of a stable relationship. The quadratic model is thus a good enough model to show that the problem is not the straight line assumption, but that there is not much association between age and spending.

The unified interaction analysis shows that the male and female age slopes are statistically different; this difference is rather negligible in practical terms. In the centered linear model, the female age slope is -$1.342 per year (95% CI: -$1.426 to -$1.258), and the age-by-male interaction is -$0.476 per year (95% CI: -$0.590 to -$0.362), giving a male slope of approximately -$1.819 per year. The full interaction model has R-squared = 0.0114, and the change in R-squared when adding the interaction is just 0.00015 (Cohen f-squared approximately 0.00015). This very large sample size thus allows one to detect a very small difference in the slope, regardless of how small it is.

The same practical conclusion is reached for the quadratic interaction model. The age-by-male interaction is -$0.569 (95% CI: -$0.684 to -$0.455), and the age-squared-by-male interaction is -$0.0511 (95% CI: -$0.0598 to -$0.0424). The R-squared for its two gender interaction terms is 0.0236, which is less than 0.001 higher than the R-squared for a quadratic model with a gender main effect (Cohen f-squared approximately 0.00061). Incorporating months engaged, purchases and support interactions brings R-squared up to 0.0617 and reduces RMSE to approximately $233.25. Support interactions have the greatest coefficient of these added variables (-$15.57 per additional interaction; 95% CI: -$15.80 to -$15.34). With that said, the extended version of the model accounts for just approximately 6.2% of the spending variation, which suggests that other variables exist that explain spending variation but were not included in the set of variables used.

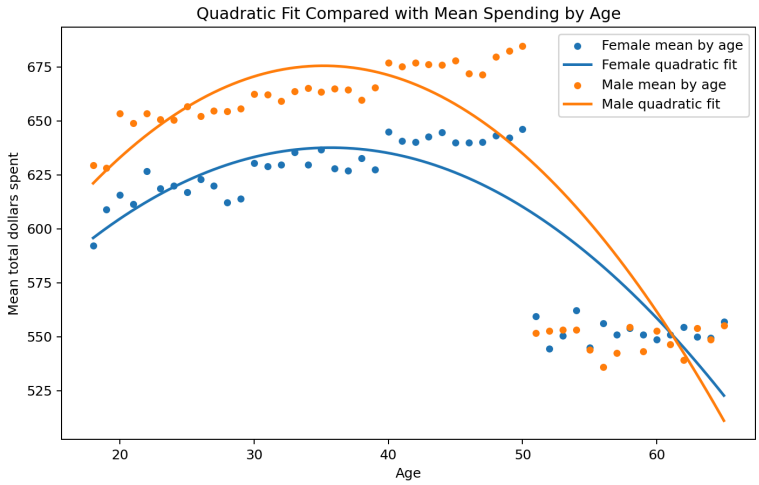


Figure 2. Quadratic fits to the mean spending at each age

DOI: 10.54254/2755-2721/2026.38785

3.3. Sampling exercise

The result obtained from Table 3 indicates that out of the first 1000 valid data, there are 514 female customers and 486 male customers. This was deemed to be sufficiently balanced under the procedure prescribed by the professor, and no further rows were needed. This conclusion should be read with a grain of caution, for the reduced sample provides a reasonable gender balance in the check, but not perfect representativeness in all variables. In a real company, a more stringent design would involve comparing distributions of age, spend, tenure and turnover or using stratified random sampling, if computation costs were significant. However, the exercise illustrates the core dilemma. The smaller sample will involve less work but the analyst has to check that the reduction has not taken away the structure required for the question under investigation.

This results, computationally, in the necessary balancing check stopping at the 1,000th record instead of growing to the full set of data, and in the main scientific analysis being based on all 440, 832 valid observations. Thus the subset is only a small efficiency statement concerning the screening task, and not a statement that 1,000 observations are adequate for all the analyses in the paper.

Table 3. Gender composition of the first 1, 000 valid records

Category	Count	Percentage
Female	514	51.4%
Male	486	48.6%
Total	1,000	100.0%

3.4. Turnover prediction results

As can be seen from Table 4, the turnovers obtained are much better than the age-spending results. Logistic regression obtains an accuracy of 82.7%, precision of 85.9%, recall of 83.0%, F1 score of 0.845 and ROC-AUC of 0.906 in the held-out test set. A significantly high AUC (>0.5) indicates that the model is able to identify actual churners more accurately than non-churners. Most customers identified as turnover customers are true positive customers (precision 0.859) and most customers who actually turn over are identified by the model (recall 0.830). Logistic regression is a useful interpretable baseline, rather than just a mathematically convenient model, due to these metrics.

Table 4. Evaluate turnover prediction for the test-set

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic regression	0.827	0.859	0.830	0.845	0.906
Random forest	0.948	1.000	0.909	0.952	0.967

The performance of the random forest is even better when tested: 94.8% accuracy, 90.9% recall, F1 score 0.952 and ROC-AUC 0.967. It has been claimed to have a precision of 1.000 at the chosen classification threshold. This improvement compared with logistic regression indicates that the relationship between turnover and the log-odds of the outcome is not linear, but has higher order interactions or relationships that are not adequately represented by a single linear log-odds equation. This trend is echoed in recent churn researches that flexible ensemble and tree-based methods can deliver good predictive results. But the above-average performance should also give the green light

DOI: 10.54254/2755-2721/2026.38785

for caution. Prior to deployment, if the company tries the model on a more recent time period, or on an outside data set and sees that no predictor was recorded after turnover or in a way that it could give away information about the target, it should be okay.

The test performance of the random forest is even better: 94.8% accuracy, 90.9% recall, F1 score 0.952 and ROC-AUC 0.967. It has a reported precision of 1.000 at the classification threshold set. The improvement from logistic regression indicates that there are nonlinear relationships or interactions in turnover which are not captured by the single linear log-odds equation. This trend mirrors recent churn studies that have demonstrated flexible ensemble and tree-based methods can offer good predictive power [3-6]. But the unusually high performance should also give reason for caution. Prior to deployment, the company ought to check the model on a subsequent time period or on an external data set and ensure that no predictor was recorded after the turnover or otherwise leaks information about the target.

3.5. Have interactions as the strongest practical signal

The most eye-catching finding is for customer-support interactions. The turnover rate is around 31% for customers who have zero to two support interactions and around 49% for customers with 3 to 4 interactions (see Figure 3). Then it increases to about 97% for 5-6 interactions and is basically 100% at the highest ranges of interaction. Another feature-importance analysis using random forest also indicates that the first feature is support interactions (43.9%), followed by total spending (30.0%) and age (22.4%). As opposed to the age-spending regression, which is widely scattered around a low regression, the grouping of the supports-interaction yields a large variation in the outcome being observed.

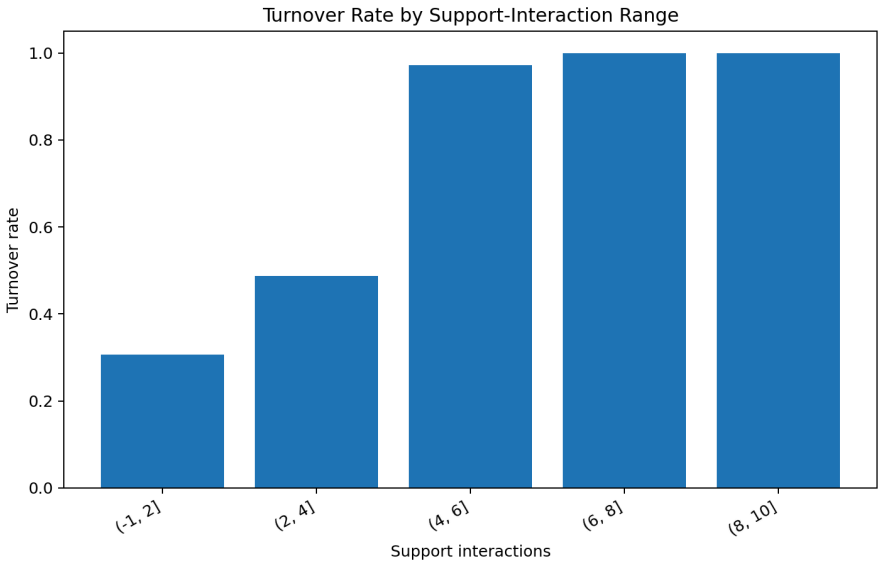


Figure 3. Turnover rate observed on the support-interaction ranges

This discovery has an obvious implication for commerce, and must be expressed with careful wording. The multiple touchpoints with support indicate that the issues are either technical in nature, billing-related, or that the customer isn't happy or there are other issues with the customer relationship. Other e-commerce studies have also shown that not providing customer support and not responding to queries are related to attrition [8]. However, this current analysis provides only

DOI: 10.54254/2755-2721/2026.38785

associations and predictive value, and not causation. But the notion that a customer contacting support leads to a 'customer loss' would be wrong. A support call from a dissatisfied customer can be a repeat of the same issue due to an initial problem, and can raise the likelihood of them leaving. The company would need a causal design like a randomized retention experiment or a well-designed quasi-experimental study to determine if a particular support program decreases churn.

3.6. Using business applications and models

A viable retention solution can be used to transform the fitted turnover model to a risk score for all current active customers. The customer could be sorted based on the probability, and the company can select a trigger level of intervention, based on the cost of the company. The threshold can be reduced if it costs a lot to lose a customer vs. sending a message to retain him or calling in a service specialist. Depending on the cost involved in outreach, the threshold may be increased, concentrating only on the highest risk cases. History of support-interaction would be particularly valuable for prioritization, in terms of predictive power and also operationally interpretable warning signal.

The model should not be used to automatically make decisions about customer treatment, but instead should be used to support human decisions. The importance of features does not define the reason for a customer to be at risk of churn and certain factors like age or gender should be given extra attention when making real business decisions. It would be good to have a sensible workflow that includes the model's risk score, the latest service context, and a record of what was done and what happened in terms of retention. These results could be fed back to the model over time to re-train and re-calibrate it. Hereby, the analysis would shift from a prediction based approach to a customer-retention system that is monitored, yet maintains the separation of prediction, explanation, and causal intervention.

3.7. Discussion: why turnover is more predictable than spending

A very significant result of this study is the divergence between the predictive powers of the spending models and of the turnover models. The R squares for the age-based spending regressions were very low. Even after incorporating a quadratic relationship, the maximum R-squared value was 0.0247 in the case of males. This means that just a small proportion of the difference in spending by customers can be attributed to their age. The multivariate demographic and behavioral models used for turnover, however, performed much better in terms of prediction, with ROC-AUC values of 0.906 (logistic regression turnover model) and 0.967 (random forest turnover model). This difference illustrates that customer information is often dependent on the target variable and information in the model.

This may be because there are many factors that affect spending behavior that are not included in the current data set. People of the same age can have very different purchasing needs, product preferences, response to promotions and financial circumstances. Thus, customers of the same age or similar can have significantly different expenditure. This is reflected in the age-spending scatterplot, which shows spending values spread out at nearly every age. A quadratic term will provide some curvature, but it will not make up for the lack of other variables that could explain purchasing behavior.

This interpretation is supported by the formal models of the interaction between gender and age that show that the age slope is different for the two genders, but that the interaction term accounts for less than one-tenth of one percentage point of the additional R-squared. The R-squared value of

DOI: 10.54254/2755-2721/2026.38785

0.0617 is achieved with the expanded spending model (spending variation from months engaged, purchases, and support interactions) but most of the spending variation is not accounted for. These effect magnitudes are more informative with regards to interpretation of the practical significance of the results than significance tests alone.

However, Turnover seems to be more associated with the behaviour information provided in the data set. Especially, there is a high correlation between support interactions and turnover. Only zero-to-two support interactions have an observed turnover rate around 31%, and it gets much higher when customers have five or more interactions. Support interactions are also the top feature by importance assigned to it by the random forest. There are other pieces to the puzzle which can be added, such as months engaged, purchases, total spending and other customer characteristics, which can be used to differentiate between higher and lower risk customers.

Another important result is the difference between logistic regression and random forest. Logistic regression models the effects of predictors as a relatively simple additive relationship to the log-odds of turnover. Random forest can model nonlinear thresholds and interactions among variables without having to prespecify the relationships. This higher ROC-AUC indicates that customer turnover may be related to a suite of behaviours not only linear, but also interacting. In general, these results demonstrate the importance of choosing the proper variables and modeling approaches, just as much as choosing a larger data set.

3.8. Interpretability, robustness, model validation

While the random forest was markedly more predictive than the logistic regression, it is not just the predictive accuracy that matters when choosing a model to use for practical purposes. Logistic regression yields a fairly interpretable link between the independent variables and the probability of turnover. This transparency can be helpful if a company requires to clarify why a customer got a high-risk score. By contrast, the random forest does not necessarily reveal such patterns or interactions between the variables, and the individual predictions are less readily interpreted. Thus, selecting one of the two models is a compromise between predictive performance and interpretability.

The high ROC-AUC of 0.967 achieved by the random forest is promising but validation of the model is important. The current train-test split considers only testing on customers that were not used to train the model, thereby minimizing the risk that the model is only tested on observations it's already seen. But people can't be sure that the same performance will persist into the future when the split is a random split. Customer behaviour, pricing and products and support procedures can change over time. This validation procedure would then be stronger, in that the model would be trained with customers at a former time and tested with customers at a subsequent time. Such temporal validation will be closer to how the model would work once deployed.

The importance of each feature should also be carefully interpreted. Although having the highest random-forest importance, support interactions don't necessarily have a direct impact on turnover. The feature importance is a measure of the utility of a variable for prediction in the fitted model. A variable may be a very strong predictor due to its connection with another process. A repeat customer can call support numerous times, and also be more likely to leave, for example, if there is an unresolved issue. In this case, support interactions serve as a symptom of customers' dissatisfaction instead of a direct driver of churn.

DOI: 10.54254/2755-2721/2026.38785

3.9. Restrictions and suggestions for future research

There are some limitations to the results of the present study. One, the spending patterns are based on the age variable. While their low R-squared values don't imply that customer spending is unpredictable, they do imply that age alone doesn't contain much information about differences in customer spending. Additional variation may be attributable to variables like income, product preference, purchase frequency, promotional exposure, and geographic location, but these variables are not included in the current data set.

Second, the turnover analysis is a prediction that is not causational. The models can recognize patterns that are indicative of customers who are more likely to leave the company, but they can't say if changing a particular predictor would make turnover lower. This is particularly significant when it comes to support interactions. A strong linkage between repeated support contacts and turnover is a valuable warning sign; however, a company should not aim to decrease support contacts just because it is a variable that predicts turnover. Future studies, however, might explore how faster resolution or better service quality or proactive follow-up action following multiple support contacts can increase retention.

An issue is also possible leakage of the targets. All predictors should be verified prior to operational use of the model to ensure that they will be available when the company makes a prediction. Test scores might look artificially high if information recorded after a customer has already determined to depart is added as a predictor. Future work should therefore specify an exact date for the prediction and only information obtained prior to that date should be taken into account. Another evaluation over time period also would be useful to see if the model remains valid as customer behaviour evolves.

Lastly, model performance should be related to business cost, just not statistically measured. The various classification thresholds generate varying numbers of false positives and false negatives. If it is much more costly to lose a customer than to reach out to a customer who would not have walked out, perhaps the company should set a lower threshold that will boost recall. In the future it might be possible to integrate the predicted turnover probabilities with the value of the customer and the cost of the act of intervention to provide a more economically meaningful definition of retention.

3.10. Practical deployment and monitoring

Ideally, in an actual customer-retention system the turnover model would need to be continually monitored, not assumed to be constant. As the prices, products, competitors and service policies change, so does the customer behavior. Because of this, a model derived from the current data set may become less and less accurate as it is applied to future customers. This precision, recall, F1 score and ROC-AUC can be periodically measured on new outcomes produced by the company and compared with the original test set results. If the percentage of decline is significant, it could mean that the model needs to be recalibrated or retrained using updated data.

The changes in the predictor distributions should also be examined in model monitoring. For instance, support interactions are the most important feature identified by the present analysis (with an importance of 43.9%) and observed turnover increases rapidly with the value of a support interaction. The distribution and interpretation of this variable may be further altered if the company alters its customer-support system. The relationship developed from the historical data could then be less representative of the future behavior. It would then be possible to get an early warning sign of model degradation by monitoring model performance and important changes of model input variables.

DOI: 10.54254/2755-2721/2026.38785

Lastly, deployment needs to check if the predictions actually have a positive business impact. The improvement of churn models to achieve an accuracy of 90% or more will not matter much if it isn't generating better customer retention. The company could measure the retention results of high-risk customers that are subjected to an intervention to similar customers that are not. This kind of testing would shift the project from being about prediction to determining whether the value of model-driven decisions can be measured. Thus, the turnover model might be incorporated into an iterative process of prediction leading to action, measurement of outcomes and feedback of new data to enhance future predictions.

4. Conclusion

In this work, models of least squares, matrix conditioning concepts, polynomial regressions, logistic regressions and random forest are applied on a large customer dataset. The theory gives the reasons why it is mathematically convenient to minimize the residual 2-norm, why a design matrix must be full column rank to have a unique least-squares coefficient vector and why the numerical use of the normal equations can be dangerous when the design matrix is ill-conditioned. These ideas are directly related to current practice in numerical linear algebra, which still relies heavily on stable solution methods for large least-squares problems.

The empirical spending analysis is calculated in a deliberately modest manner. The age slopes are negative for women and men and only account for less than one per cent of variation in spending. Minor curvature is manifested in the quadratic models, and the fit is improved, but even the best model accounts for only 2.47% of the variation. Hence, age should not be seen as a strong indicator of personal expenditure. The separate first-1, 000 sampling exercise provides evidence of a simple computational-cost strategy and shows a 51.4/48.6 female/male split which is adequate for the professor stipulated screening rule.

A single gender difference in the age slope was confirmed through a unified interaction analysis, which only contributes 0.00015 to the R-squared of the linear model. The gender interactions are also small in terms of their contribution to the explanation of the model. The inclusion of the various engagement, purchase, and support factors boosts the R-squared for the spending model to 0.0617, but most of the variation in spending remains unaccounted for. These results focus on effect size and not statistical significance in a large sample of participants as proof of practical significance.

Turnover prediction is the most robust conclusion. Demographic and behavioral variables combined come to ROC-AUC 0.906 with logistic regression and 0.967 with random forest. The most obvious practical indicator of risk is support interaction, as the more interactions that are observed, the more turnover that is seen as the interaction range increases. The key finding of the project is this contrast: variables which account for little in terms of customer expenditure can prove to be valuable in terms of customer retention. Future work should test and validate the models on later time periods, test for target leakage, add richer transaction and complaint-resolution data and causally test retention interventions. These extensions would boost the accuracy of the predictions and the company's ability to translate prediction into good customer-retention decisions.

References

[1] Wathen, A. J. (2025). Least squares and the not-normal equations. *SIAM Review, 67*(4), 865–872. https://doi.org/10.1137/23M161851X

[2] Meier, M., Nakatsukasa, Y., Townsend, A., & Webb, M. (2024). Are sketch-and-precondition least squares solvers numerically stable? *SIAM Journal on Matrix Analysis and Applications, 45*(2), 905–929. https://doi.org/10.1137/23M1551973

DOI: 10.54254/2755-2721/2026.38785

[3] Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks, 4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005
[4] Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. *Journal of Big Data, 10*, article 41. https://doi.org/10.1186/s40537-023-00721-8
[5] Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis. *Technologies, 11*(6), article 167. https://doi.org/10.3390/technologies11060167
[6] Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms, 17*(6), article 231. https://doi.org/10.3390/a17060231
[7] Mena, G., Coussement, K., De Bock, K. W., De Caigny, A., & Lessmann, S. (2024). Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. *Annals of Operations Research, 339*, 765–787. https://doi.org/10.1007/s10479-023-05259-9
[8] El Koufi, N., & Belangour, A. (2024). Toward a recommender system for assisting customers at risk of churning in e-commerce platforms based on a combination of Social Network Analysis (SNA) and deep learning. *Journal of Open Innovation: Technology, Market, and Complexity, 10*(4), article 100425. https://doi.org/10.1016/j.joitmc.2024.100425
[9] Zhang, H., & Zhang, W. (2024). Application of GWO-attention-ConvLSTM model in customer churn prediction and satisfaction analysis in customer relationship management. *Heliyon, 10*(17), article e37229. https://doi.org/10.1016/j.heliyon.2024.e37229
[10] Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (4th ed.). Johns Hopkins University Press.
[11] Trefethen, L. N., & Bau, D., III. (1997). *Numerical linear algebra*. Society for Industrial and Applied Mathematics.
[12] Higham, N. J. (2002). *Accuracy and stability of numerical algorithms* (2nd ed.). Society for Industrial and Applied Mathematics.
[13] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.