Predicting Employee Turnover in the Financial Company: A Comparative Study of CatBoost and XGBoost Models

Ziqing Yin *, Baojun Hu, Shuhan Chen

Posted Date: 2 October 2024

doi: 10.20944/preprints202410.0072.v1

Keywords: Employee turnover; CatBoost; XGBoost; Machine learning

 doi:10.20944/preprints202410.0072.v1

Article

Predicting Employee Turnover in the Financial Company: A Comparative Study of CatBoost and XGBoost Models

Ziqing Yin [1,*], Baojun Hu [2] and Shuhan Chen [3]

[1] University of Melbourne, Australia
[2] Warrington college of business, university of florida, gainesville, florida, USA; corinthpersonal@163.com
[3] Guanghua School of Management, Peking University, Beijing, China; jincyisme@163.com
* Correspondence: Correspondence: 1954351791@qq.com

Abstract. Employee turnover is a significant issue for financial institutions, impacting productivity, increasing recruitment costs, and disrupting critical operations. In this project, this study aimed to predict employee turnover using a dataset containing attributes such as employee satisfaction, performance, and tenure. By framing the task as a binary classification problem, this study employed CatBoost and XGBoost, two advanced regression-based algorithms, to develop predictive models. This paper's analysis demonstrated that CatBoost outperformed XGBoost across all evaluation metrics, including MAE, MSE, RMSE, and R^2, making it the more effective model for predicting turnover in the financial sector. The study highlights key factors contributing to employee attrition, such as job satisfaction, tenure, and promotion opportunities, offering actionable insights for retention strategies. Additionally, by predicting probabilities rather than binary outcomes, this study aims to make more detailed decisions about employee retention. This research provides valuable tools for financial institutions to mitigate the risk of turnover, retain critical talent, and ensure operational continuity.

Keywords: Employee turnover; CatBoost; XGBoost; Machine learning

1. Introduction

Employee turnover poses a significant challenge for financial institutions, directly affecting productivity and driving up costs related to recruitment, hiring, and training new staff. In the highly competitive and regulated environment of the financial sector, employee attrition can disrupt critical operations, impact client relationships, and hinder long-term growth. Consequently, predicting employee departure becomes essential for retaining key talent, safeguarding organizational knowledge, and ensuring operational stability.

In this project, the goal is to predict the likelihood of employee turnover using a dataset with features such as employee satisfaction, performance, and tenure. The financial industry often employs a highly skilled workforce, making it particularly important to anticipate which employees might leave and proactively address retention risks. The task is framed as a binary classification problem, predicting whether an employee will stay or exit the organization.

To address this challenge, this study have selected advanced regression-based algorithms like CatBoost and XGBoost, which are well-suited to handling the unique complexities of financial datasets. CatBoost excels in processing categorical data, commonly found in finance, and includes mechanisms to prevent overfitting, while XGBoost offers superior computational speed, regularization options, and efficient handling of imbalanced datasets. By leveraging these models, this study aim to identify the best strategy for predicting turnover within financial firms, ultimately providing actionable insights for improving employee retention and maintaining the continuity of critical financial services.

2. Related Work

 doi:10.20944/preprints202410.0072.v1

Employee turnover has been a widely researched topic in organizational studies, with various factors influencing employee retention and attrition. Previous research has focused on several dimensions, including human capital, social networks, organizational innovation, social insurance, and specific human capital, as well as risk analysis and control.

Liu et al. explored the impact of human capital and social networks on organizational innovation using online resume data, revealing that the quality and diversity of an organization's human capital significantly influence innovation capabilities. They emphasized that a well-structured social network within an organization can foster information exchange and collaboration, ultimately reducing employee turnover by enhancing job satisfaction and engagement [1].

Zhang and Yang provided an analysis of the risks associated with employee mobility, noting that turnover can be both a risk and an opportunity for organizations. They highlighted the need for companies to establish effective risk control mechanisms, such as developing robust retention strategies and implementing succession planning to mitigate the adverse effects of turnover on organizational stability and performance [2].

Huang et al. examined the dynamics of employee turnover in small and micro enterprises in China, identifying key factors that affect employee mobility, such as compensation, job satisfaction, and organizational culture. Their findings suggest that turnover can have a profound impact on the performance and survival of small businesses, where the loss of a single employee can disrupt operations significantly [3].

Yang and Lian studied the effect of social insurance on employee resignation rates in China using a difference-in-differences model. Their research indicates that providing comprehensive social insurance can lower turnover rates, as it enhances job security and employee welfare. This finding underscores the importance of employee benefits and social security in retention strategies [4].

Hu and Lu discussed the concept of firm-specific human capital and its relationship with employee mobility. They argued that employees who possess skills and knowledge specific to a particular firm are less likely to leave, as their value outside the organization may be limited. This theory aligns with the view that investing in employee development tailored to the organization's needs can be a strategic move to reduce turnover [5].

Lastly, Wang proposed a decision model for employee turnover and retention, which integrates various factors such as personal characteristics, job satisfaction, organizational commitment, and external job market conditions. The model offers a comprehensive framework for understanding the complex decision-making processes behind employee turnover, providing valuable insights for developing targeted retention policies [6].

3. Data

3.1. Variable Introduction

The data set used in this project contains key information about employees of financial companies and various attributes that may affect their decision to leave or stay in the company. Each variable captures the specific characteristics related to employee satisfaction, performance and tenure, which provides valuable insights for predicting employee turnover rate. The following is a summary of dataset variables:

Table 1. Dataset Description.

Variable	Description
satisfaction_level	The level of satisfaction of the employee
last_evaluation	The score of the last evaluation of the employee
number_project	The number of projects the employee has worked on
average_montly_hours	The average monthly hours worked by the employee
time_spend_company	The number of years the employee has spent at the company
Work_accident	Whether the employee had a work accident (1 = yes, 0 = no)
left	Whether the employee has left the company (1 = yes, 0 = no)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2024 doi:10.20944/preprints202410.0072.v1

promotion_last_5years	Whether the employee had a promotion in the last 5 years (1 = yes, 0 = no)
sales	The department the employee works in
salary	The salary level of the employee (low, medium, high)

The dataset provides a rich mix of continuous and categorical variables that capture multiple dimensions of an employee's experience at the company. Key features such as satisfaction_level, last_evaluation, and average_monthly_hours reflect the employee's job engagement and work-life balance, while variables like time_spent_company and number_project help to quantify the employee's tenure and workload. Additionally, categorical variables like sales (which refers to the department) and salary (indicating salary level) offer insights into the potential influence of organizational structure and compensation on employee decisions. The binary outcomes of Work_accident, promotion_last_5years, and left provide important context for evaluating employee turnover risks.

By analyzing these features, the model aims to predict the likelihood of an employee leaving, offering organizations the potential to design targeted retention strategies.

3.2. Data Visualization

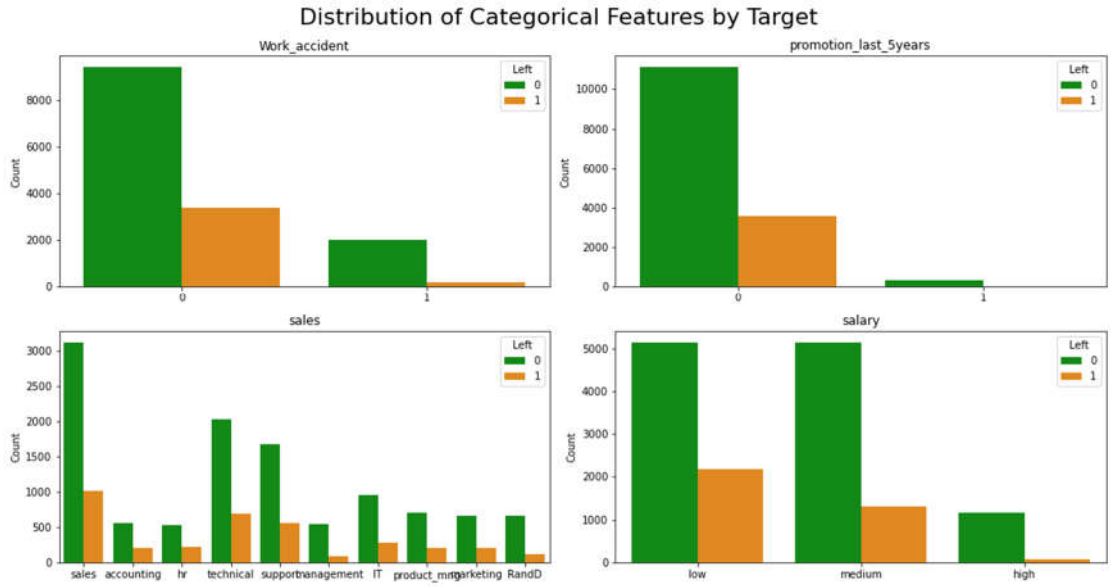

Figure 1. Distribution of Categorical Features by Target.

The Work_accident plot shows that employees who had a work accident are less likely to leave the company. The promotion_last_5years plot shows that employees who have not received a promotion in the last 5 years are more likely to leave the company. The sales plot shows the distribution of employees who left or did not leave the company across different departments. It seems that the sales, technical, and support departments have the highest number of employees who left the company. The salary plot shows that employees with a low salary are more likely to leave the company than those with a medium or high salary.

3.3. Descriptive Analysis

Based on the descriptive statistics, there are noticeable differences in employee satisfaction, evaluation scores, project numbers, average monthly working hours, and tenure. Firstly, the average employee satisfaction is 0.61 with a standard deviation of 0.25, indicating variability in satisfaction levels across employees. The minimum value is 0.09, and the maximum is 1.00, showing that some employees have very low satisfaction, while others are highly satisfied. Overall, most employees exhibit a moderate to high level of satisfaction. Secondly, the average score for the last evaluation is

 doi:10.20944/preprints202410.0072.v1

0.72, with a standard deviation of 0.17, suggesting a relatively concentrated distribution. One-quarter of employees scored below 0.56, while over half scored above 0.72, indicating that most employees this studyre rated as performing well.

Table 2. Descriptive Statistical Analysis.

	count	mean	std	min	0.25	0.5	0.75	max
satisfaction_level	14999	0.61	0.25	0.09	0.44	0.64	0.82	1
last_evaluation	14999	0.72	0.17	0.36	0.56	0.72	0.87	1
number_project	14999	3.80	1.23	2.00	3.00	4.00	5.00	7
average_montly_hours	14999	201.05	49.94	96.00	156.00	200.00	245.00	310
time_spend_company	14999	3.50	1.46	2.00	3.00	3.00	4.00	10

The average number of projects employees participated in is 3.8, with a standard deviation of 1.23, reflecting substantial differences in project involvement. The minimum number of projects is 2, while the maximum is 7, with most employees handling between 3 and 5 projects. Additionally, employees work an average of 201.05 hours per month, with a standard deviation of 49.94 hours, revealing significant differences in working hours. The minimum monthly hours worked is 96, while the maximum is 310. Most employees work between 156 and 245 hours per month, indicating a generally high workload. The average tenure of employees at the company is 3.5 years, with a standard deviation of 1.46 years. Both the 25th and 50th percentiles are at 3 years, showing that most employees have been with the company for 3 to 4 years, with the longest tenure reaching 10 years. This suggests that while many employees stay for a few years, some have much longer tenures.

In summary, there are clear individual differences in employee satisfaction, performance, project participation, and working hours. Most employees participate in a moderate number of projects, have relatively high evaluation scores, and exhibit moderate to high satisfaction. The tenure of employees is mainly concentrated between 3 and 4 years, which may indicate some level of employee turnover, despite the relatively long working hours.

4. Modeling

CatBoost and XGBoost are two powerful machine learning algorithms used in both regression and classification tasks. Both algorithms belong to the family of gradient boosting models, but each has unique characteristics and advantages.

This study are going to use regression algorithms instead of classification algorithms in this project for three main reasons:

Probabilistic Interpretation: Regression algorithms can predict a continuous output which can be interpreted as the probability of a certain event. In this case, the output can be interpreted as the probability that an employee will leave the company. This information can be more informative than just a binary output and can help in understanding how 'at risk' each employee is of leaving.

Threshold Calibration: By predicting probabilities, this study can adjust the threshold for classifying an observation as 0 or 1. For example, this study might classify all employees with a predicted probability of leaving greater than 0.5 as 'will leave'. However, this study can adjust this threshold to be more conservative or more liberal depending on the cost of false positives and false negatives. For example, if it is more costly to incorrectly predict that an employee will stay when they actually leave, this study might lower the threshold to 0.3 to identify more employees at risk of leaving.

Imbalanced Data: The target variable, left, is imbalanced with a larger proportion of employees who did not leave the company. This can sometimes lead to poor performance for classification algorithms because they have a bias towards the majority class. Regression algorithms do not have this bias and can sometimes perform better on imbalanced data.

 doi:10.20944/preprints202410.0072.v1

Using regression algorithms will allow for a more nuanced understanding of the risk of each employee leaving, allow for threshold calibration, and might perform better on imbalanced data.

4.1. CatBoost

CatBoost, a machine learning algorithm developed by Yandex, is an implementation of gradient boosting, optimized specifically for categorical features (hence the name CatBoost, Cat stands for categorical) [7]. The key advantages of CatBoost include:

Handling of Categorical Features: Unlike many other machine learning models that require extensive preprocessing for categorical variables, CatBoost is capable of processing categorical variables automatically, which simplifies data preparation and often leads to superior results.

Avoids Overfitting: CatBoost uses a special algorithm to avoid overfitting, which is a common problem with gradient boosting models. The algorithm, called Ordered Boosting, involves creating a new permutation of the training dataset for each iteration and then using accumulated statistics from the preceding part of the dataset to calculate the values for a given categorical feature.

Speed and Performance: CatBoost provides superior performance in many scenarios, offering competitive results with less tuning required compared to other models. It also provides a fast prediction speed [8].

4.2. XGBoost

XGBoost, short for Extreme Gradient Boosting, is an open-source machine learning algorithm developed by Tianqi Chen. It also belongs to the family of gradient boosting models and is known for its speed, efficiency, and high performance [9]. The main characteristics of XGBoost include:

Regularization: XGBoost offers a more regularized form of gradient boosting, which helps to prevent overfitting. It includes additional parameters for tuning the model and allows for L1 (Lasso Regression) and L2 (Ridge Regression) regularization.

Handling of Sparse Data: XGBoost is capable of handling sparse data or missing values. It uses a specific method to find the best split at each node, considering both the features that have a missing value and those that do not [10].

Parallel Processing: XGBoost features parallel processing that makes it faster than other gradient boosting algorithms. It also includes a range of other features for model tuning, cross-validation, and handling missing values.

4.3. Comparative Analysis of Models

In comparing the performance of the CatBoost and XGBoost models based on the provided evaluation metrics (MAE, MSE, RMSE, and R^2), CatBoost outperforms XGBoost across all measures.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2024 doi:10.20944/preprints202410.0072.v1

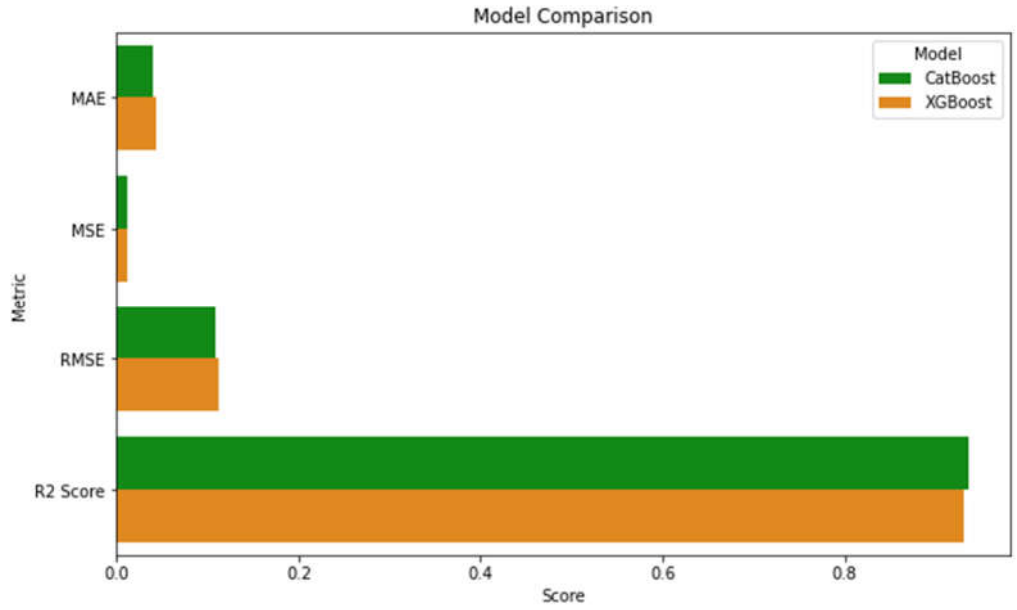

Figure 2. Model Comparison.

MAE (Mean Absolute Error): CatBoost has a lower MAE of 0.03919 compared to XGBoost's 0.043599. This means that on average, CatBoost's predictions deviate from the actual values by a smaller margin, indicating better accuracy in predictions.

MSE (Mean Squared Error): CatBoost also achieves a lower MSE of 0.011617, while XGBoost's MSE is slightly higher at 0.01261. Since MSE penalizes larger errors more severely, this suggests that CatBoost handles outliers or larger errors more effectively than XGBoost.

Table 3. Comparison Table of Model Results.

Model	MAE	MSE	RMSE	R2 Score
CatBoost	0.039190	0.011617	0.107781	0.935945
XGBoost	0.043599	0.012610	0.112295	0.930467

RMSE (Root Mean Squared Error): Similar to MSE, the RMSE for CatBoost (0.107781) is lower than for XGBoost (0.112295). RMSE is a more interpretable measure of average prediction error in the same units as the target variable, further demonstrating CatBoost's superior predictive performance.

R^2 Score (Coefficient of Determination): CatBoost achieves an R^2 score of 0.935945, while XGBoost has a slightly lower R^2 score of 0.930467. The R^2 score indicates how well the model explains the variance in the data. Both models perform well, but CatBoost explains a slightly higher proportion of the variance, making it the better model in this comparison.

In summary, CatBoost demonstrates better performance across all metrics compared to XGBoost, indicating that it is the more accurate and reliable model for this regression task.

5. Conclusions

This project demonstrates the successful application of advanced machine learning algorithms, specifically CatBoost and XGBoost, in predicting employee turnover within the financial sector. By leveraging a rich dataset of employee attributes, including satisfaction levels, performance metrics, and tenure, this study is able to develop predictive models that help organizations better understand which employees are at higher risk of leaving. Such insights are crucial for improving retention strategies and ensuring the continuity of critical financial services.

The analysis of this paper shows that CatBoost is superior to XGBoost in all evaluation indexes, including MAE, MSE, RMSE and R^2, which makes it the best model to predict the flow of people in

 doi:10.20944/preprints202410.0072.v1

this situation. CatBoost's ability to efficiently handle classified variables and avoid over-fitting has played a key role in its success, while XGBoost has also produced reliable results in regularization and processing of unbalanced data.

The findings emphasize the importance of addressing employee satisfaction, performance, and tenure as critical factors in retention efforts. Furthermore, the project underscores the value of using probabilistic interpretations of model outputs, allowing for nuanced, data-driven decisions regarding employee retention strategies. In a highly competitive industry, the ability to predict and mitigate employee turnover offers a significant strategic advantage, helping organizations to retain key talent and maintain operational stability.

Future work may focus on refining the models by incorporating additional features such as external market data or personal characteristics, further improve the prediction ability of the algorithm. Additionally, applying these models to other industries with similar employee retention challenges could offer broader insights into turnover patterns across different sectors.

References

1. Liu Shanshi, Sun Bo, Ge Chunmian, et al. Social network of human capital and enterprise innovation-an empirical study based on online resume data [J]. Management World, 2017, (07): 88-98+119+188. doi: 10.19744/j.cnki.11.
2. Zhang Yali, Yang Naiding. Risk analysis and control of personnel flow [J]. Science and Science and Technology Management, 2000,(09):42-44.
3. Huang Yuhong, Yi Daichun, Jie Mengyin. Analysis of the current situation, role and influencing factors of employee mobility in small and micro enterprises in China [J]. Management World, 2016, (12): 77-89. doi: 10.9744/j.cnki.11-1235/F.2016.
4. Yang Yinan, Lian Yujun. Can social insurance reduce employee resignation rate? -estimation of the double difference model of comprehensive social survey in China [J]. Economic Management, 2015,37 (01): 168-179.doi: 10.19616/j.cnki.bmj.2015.01.019.
5. Hu Haozhi, Lu Xianxiang. Enterprise-specific human capital and employee mobility [J]. Financial Research, 2010,(06):86-92.
6. Wang Chunxiu. Decision-making model of employee leaving and staying [J]. Shopping Mall Modernization, 2010,(35):157-159.
7. Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 94.
8. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.
9. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
10. Ogunleye A, Wang Q G. XGBoost model for chronic kidney disease diagnosis[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2019, 17(6): 2131-2140.