BY USING DATA ANALYSIS AND MACHINE LEARNING TECHNIQUES FOR BANK TURNOVER PREDICTION

Meliboev Azizjon
Department of Digital technologies and mathematics, Kokand University

Abstract: Bank turnover prediction entails utilizing data analysis and machine learning methods to anticipate the probability of customers leaving or churning from a bank. This predictive modeling approach relies on historical customer data, including various attributes like transaction history, account details, demographics, and patterns of customer behavior. By employing advanced analytical techniques and machine learning algorithms, the objective is to forecast and address customer churn by identifying significant contributors to turnover. In this study, we conduct data preprocessing, feature engineering, and exploratory analysis to glean insights from the dataset. Furthermore, we employ machine learning algorithms such as Random Forests, to develop the predictive model. These models are then validated and refined using methods like hyperparameter tuning, along with evaluation metrics such as accuracy of 86%.

Key words: — Bank turnover, Machine learning, Data analysis, Random forest

Introduction:

In the dynamic and competitive field of banking, retaining employees is essential to uphold a proficient workforce, maintain operational effectiveness, and cultivate client satisfaction. Nonetheless, the issue of employee turnover, also referred to as churn, persists as a significant challenge, resulting in notable financial and reputational setbacks. To effectively tackle this challenge, banks are increasingly relying on data analysis and machine learning methods to anticipate which employees may be inclined to leave. By harnessing the capabilities of data analytics, banks can extract valuable insights regarding employee demographics, work history, performance appraisals, compensation records, and other pertinent variables that influence turnover. Through machine learning algorithms trained on comprehensive datasets, banks can discern patterns and correlations indicative of a heightened probability of employee departure. This proactive approach to managing turnover provides banks with numerous advantages.

Utilizing data analysis and machine learning methods to predict bank turnover offers a revolutionary solution to the difficulties of retaining employees. Leveraging data-driven insights enables banks to preemptively recognize employees who may be considering leaving, allowing them to deploy tailored retention tactics and improve overall employee involvement and contentment. This proactive method not only alleviates the financial strain of turnover but also cultivates a more robust, customer-focused banking workforce.

In recent times, the emergence of sophisticated machine learning algorithms within the field of computer science has spurred the development of robust quantitative methods for extracting valuable insights from industry data. Specifically, supervised machine learning techniques, which involve computers learning from in-depth analyses of extensive and well-labeled historical datasets, have exhibited their proficiency in extracting meaningful insights across various domains. Scholars have explored numerous machine learning methodologies to enhance human resource (HR) management outcomes. Various supervised machine learning algorithms, including Decision Trees, Random Forests, Gradient Boosting Trees, Logistic Regression, and Support Vector Machines, have been detailed, demonstrated, and assessed for their efficacy in predicting employee turnover.

The concept of "Churn Modeling," also known as customer attrition analysis, holds significant importance in customer relationship management and predictive analytics. Several research works delve into different aspects of churn modeling. Notably, a suggested model in one study exhibited notably higher accuracy compared to Random Forest, particularly in churn prediction. Artificial Neural Networks (ANN) displayed the highest accuracy among supervised machine learning techniques, whereas Decision Trees (DT) showed comparatively lower accuracy. These findings underscore the advantage of boosting variants over simpler models and bagging methods in terms of predictive performance.

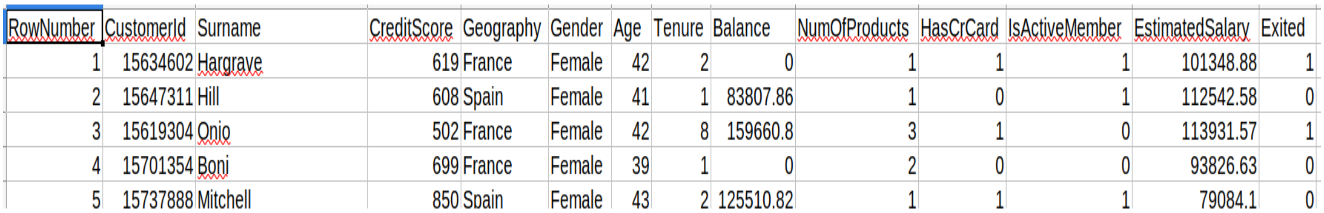
In another study, the focus shifted towards utilizing neural networks to forecast customer churn within the banking sector, which is crucial for customer retention efforts. The study aimed to present a case study illustrating the application of data mining techniques, particularly neural networks, for extracting insights from banking sector databases. The findings indicated that clients engaging with a greater number of bank services tend to exhibit higher loyalty, suggesting a strategy for the bank to concentrate on clients utilizing fewer than three products and tailor offerings to their specific needs.

Additionally, another exploration investigated various machine learning algorithms in building churn models to aid telecom operators in predicting potential customer churn. Experimental results highlighted the superiority of utilizing Random Forest in conjunction with SMOTE-ENN in terms of F1-score compared to other approaches.

Furthermore, a study aimed to predict customer churn using a range of models, employing machine learning techniques such as Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machines, AdaBoost, Multi-Layer Sensors, and Naive Bayes on relevant datasets. The analysis revealed that the Random Forest method proved most effective in predicting customer attrition for both datasets.

In this work, the objective is to demonstrate data analytics and apply machine learning algorithms such as Random Forest on the pertinent dataset. This dataset [8] that we used, comprises information about customers of a bank sourced from Kaggle, with the focal point being a binary variable indicating whether a customer has opted to terminate their account or if they have chosen to maintain their association with the bank. The dataset contains information of 10,000 bank customers with total of 11 features which are ’CreditScore’, ’Geography’, ’Gender’, ’Age’, ’Tenure’, ’Balance’, ’NumOfProducts’, ’HasCrCard’, ’I sActiveMember’, ’EstimatedSalary’, ’Exited’. This variable serves as a pivotal indicator of customer retention, a critical metric for financial institutions striving to uphold long-term relationships with their clienteleas shown Figure 1.

Fig. 1. 5 head of data description

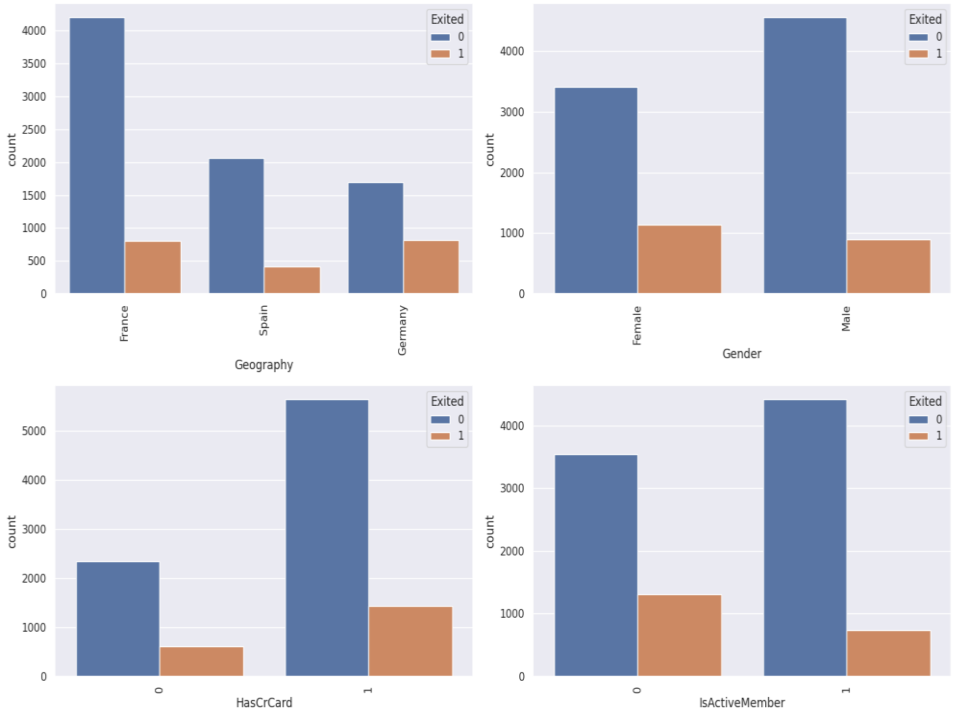


RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0

The dataset encompasses a diverse range of attributes pertaining to these customers, allowing for a comprehensive analysis of factors influencing their decision to stay or leave. This information provides a valuable foundation for leveraging advanced analytical techniques and machine learning algorithms to develop a predictive model for customer churn.

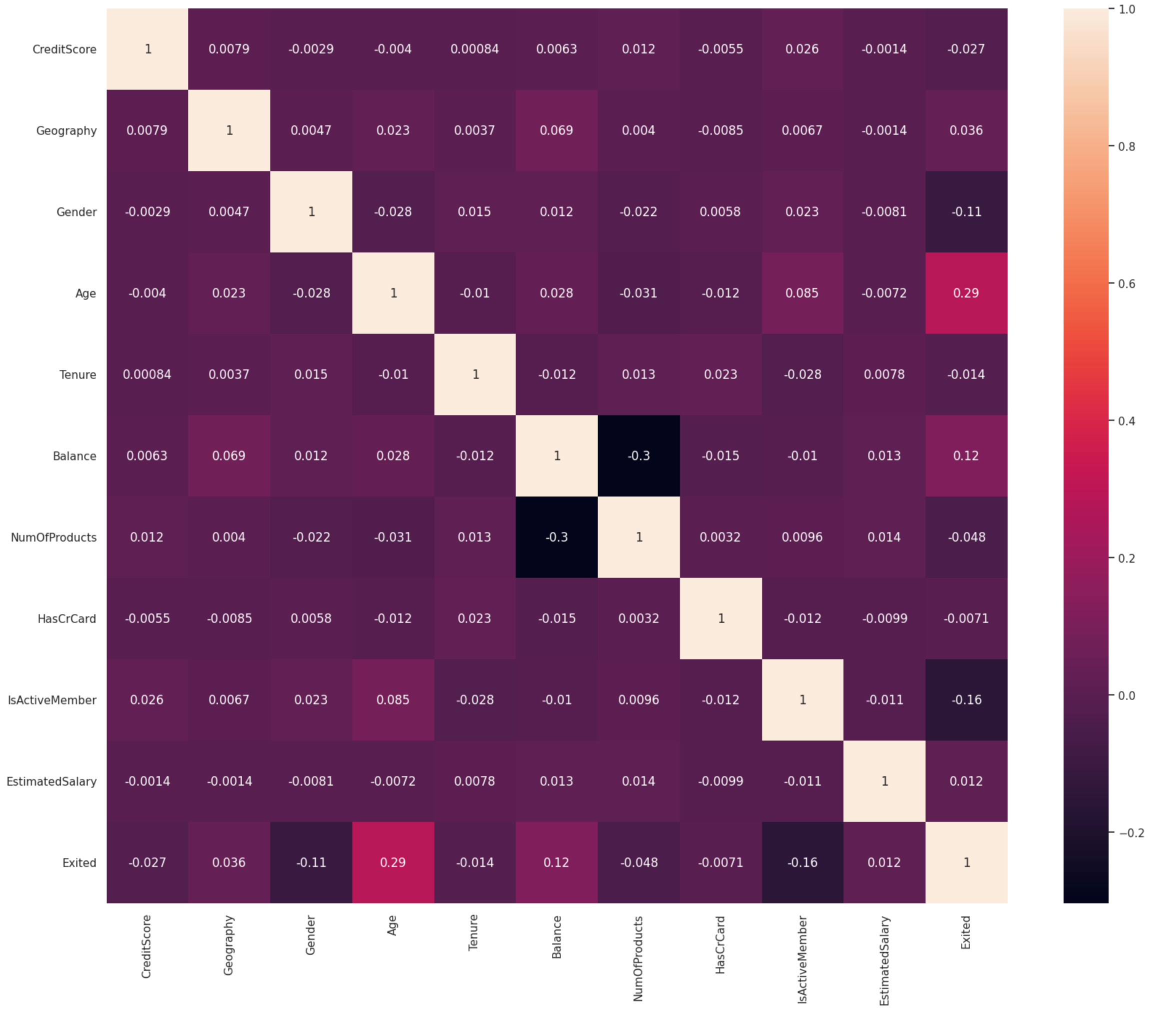
In this part we have used label encoding process on categorical columns in our data. That need to be converted to numerical values for machine learning algorithms that require numerical input. We set categorical variables as a list which contain the names of categorical variables that we want to analyze. These could be columns that are important features in our dataset representing categories which are 'Geography' has values like 'France', 'Spain', 'Germany'; 'Gender' feature has value 'Female', 'Male'; ''HasCrCard' that defines has credit card has values 'Yes', 'Not'; "IsActiveMember" has also values 'Yes', 'Not'. The String values transformed that 'Yes' mentioned as 1, 'Not' is mentioned as 0. We represent plot of visualizing the distribution of categorical variables in our dataset that is specifically in the context of customer as shown Figure 2.

Fig. 2. Bar feature distribution.



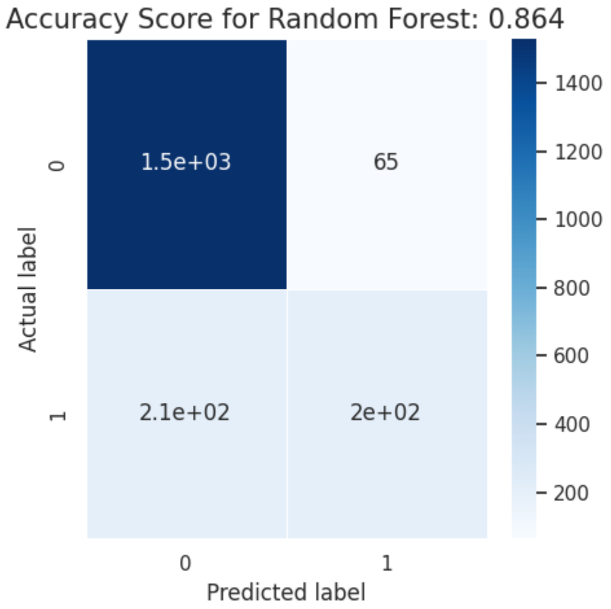
This section lays the groundwork for the experiments detailed in the subsequent section. We introduce essential metrics for evaluating model performance, including Model Evaluation and the Confusion Matrix. The Confusion Matrix offers a visual depiction of a model's effectiveness, distinguishing between accurately classified positive cases (True Positives) and incorrectly rejected instances (False Positives), which signify the number of benign conditions erroneously identified as malignant. Additionally, it discerns between correctly identified negative instances (True Negatives) and incorrectly identified instances (False Negatives), indicating the number of malignant tumors mistakenly identified as benign. Further elaboration on these concepts can be found in our prior research [11]. A table is provided to summarize the definitions within the Confusion Matrix. Our models were assessed using the following metrics:

Fig. 3. Heatmap of the correlation matrix.



We applied two reliable machine learning algorithms which are Random forest. According to the results, the Random Forest model achieved its highest performance after tuning hyperparameters. Figure 4.

Fig. 4. Accuracy score and Confusion matrix of Random forest model.



In summary, the application of data analysis and machine learning methods for predicting bank turnover has become a potent asset for banks to actively handle employee retention challenges and minimize the negative impact of employee turnover. Employing sophisticated algorithms like Random Forest and K-Nearest Neighbors (KNN) enables us to glean valuable insights into employee behavior, identify individuals at risk of leaving, and execute tailored retention tactics. Based on the outcomes of our models, Random Forests demonstrate consistent performance.

References:

1. Q.A. Al-Radaideh, E. AlNagi, ”Using data mining techniques to build a classification model for predicting employees performance,” International Journal of Advanced Computer Science and Applications vol. 3, no. 2, 144–151 2012.

2. H.Y. Chang, ”Employee turnover: a novel prediction solution with effective feature selection,” The WSEAS Transactions on Information Science and Applications vol. 6, pp. 417–426, 2009.

3. Y. Zhao, M.K. Hryniewicki, F. Cheng, B. Fu and X. Zhu, ”Employee turnover prediction with machine learning: A reliable approach.” Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) vol. 2, pp. 737-758, 2019.

4. S. Khodabandehlou and M., Zivari Rahman, ”Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior,” Journal of Systems and Information Technology, n 1/2, pp. 65-93, 2017.

5. A, Bilal Zorić,. ”Predicting customer churn in banking industry using neural networks.” Interdisciplinary Description of Complex Systems: INDECS vol. 14, no. 2, pp. 116-124, 2016.

6. R., Srinivasan, D. Rajeswari, and G. Elangovan. ”Customer Churn Prediction Using Machine Learning Approaches.” International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) pp. 1-6, 2023

7. H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, ”Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis,” International Conference on Computer Science and Engineering (UBMK), pp. 139- 144, 2021.

8. Data: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling

9. F., Pedregosa, G., Varoquaux, A., Gramfort, V., Michel, B., Thirion, O., Grisel, É., Duchesnay, ”Scikit-learn: Machine learning in Python”. the Journal of machine Learning research, no 12, pp. 2825-2830, 2011.

10. T., Akiba, S., Sano, T., Yanase, T. Ohta and M., Koyama, ”Optuna: A next-generation hyperparameter optimization framework.” Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining pp. 2623-2631. 2019

11. A., Meliboev, J., Alikhanov and W., Kim. ”Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets”. Electronics, 11(4), p.515. 2022.

12. Z.H. Hoo, J. Candlish, and D. Teare. ”What is an ROC curve?.” Emergency Medicine Journal, 34(6), pp. 357-359. 2017.

13. Sh. Fazilov, O. Mirzaev, and Sh. Kakharov. ”Building a Local Classifier for Component-Based Face Recognition.” International Conference on Intelligent Human Computer Interaction, pp. 177-187, 2022.

14. Abdullajonov, D., & Qosimova, G. (2022). O ‘ZBEKISTONDA MASOFAVIY TA ‘LIMINI TASHKIL ETISH VA RIVOJLANTIRISH. *Евразийский журнал академических исследований*, *2*(13), 1402-1407.

15. Abdullajonov, D. (2021). RAQAMLI TEXNOLOGIYALAR ORQALI YANGI O’ZBEKISTONNING IQTISODIYOTINI RIVOJLANTIRISH, RAQAMLI IQTISODIYOTNING ISTIQBOLLARI. *Экономика и социум*, (12-1 (91)), 28-33.

16. Toxirjon, U. (2024). BOSHLANGICH SINFLARDA O ‘QISHNI YETKAZIB OLISHGA QIYNALAYOTGAN O ‘QUVCHILAR BILAN ISHLASHDA INTERFAOL USULLARDAN FOYDALANISH. Integration of Economy and Education in the 21st century, 2(2), 9-13.

17. Toxirjon, U. (2024). XALQARO O ‘QISH SAVODXONLIGINI O ‘RGANISH (PIRLS). Integration of Economy and Education in the 21st century, 2(2), 14-17.