Customer Churn Prediction in Banking Sector by Using ML Techniques

Meetkumar Dilipbhai Kanani[1], kananimeet34@gmail.com

Bhargav Dipakkumar Amin[1], bhargavamin2013@yahoo.in

Dr. Akil Z. Surti[2], dr.akilzsurti@gmail.com

Abstract:

Customer churn is a major challenge for businesses across various industries. It refers to the situation where customers end their relationship with a company or brand, which can cause significant revenue loss. To mitigate this loss and implement proactive retention strategies, accurate prediction of churn is crucial. In this paper, we conduct a comprehensive study of machine learning techniques for customer churn prediction. We use real-world datasets from the banking sector to compare the performance of various algorithms and provide insights into feature importance. Our finding indicates that random forests are highly effective in predicting churn. This contributes to the advancement of predictive modeling in customer relationship management.

KEYWORDS: Customer churn, Predictive modeling, Machine learning, Classification algorithms, Feature importance

I. INTRODUCTION

Customer churn, which refers to the loss of customers, is a significant challenge for businesses across various industries. When customers discontinue their relationship with a company or brand, it leads to lost revenue and decreased profitability. Hence, accurate prediction of churn is crucial for businesses to implement targeted retention strategies and maintain customer satisfaction. With the proliferation of data and advancements in machine learning techniques, predictive modeling has emerged as a powerful tool for identifying customers at risk of churn. To this end, we investigate the application of various machine learning algorithms to predict customer churn. Our goal is to provide insights into their comparative performance and identify key drivers of churn. Predicting customer churn has become a focal point for businesses seeking to optimize customer retention strategies. Traditional approaches to churn prediction often relied on simplistic models and manual analysis of customer behavior. However, with the advent of big data and advancements in machine learning techniques, businesses now have access to more sophisticated tools for churn prediction.

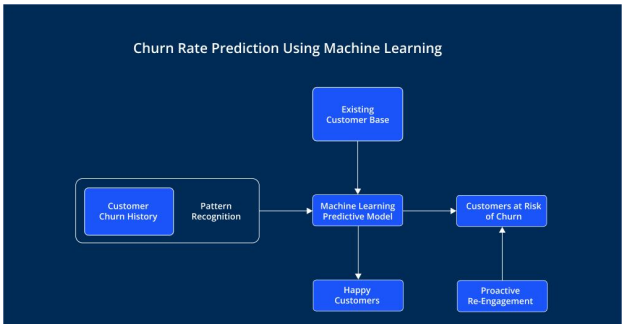

Fig 1. overview of churn prediction model

II. LITERATURE REVIEW

Research on customer churn prediction has used various methodologies and approaches. Traditional statistical techniques like logistic regression and survival analysis have been widely used due to their interpretability and ease of implementation. These methods rely on predefined rules or assumptions about the relationship between predictor variables and churn.

Recently, there has been a shift towards the use of machine learning algorithms for churn prediction. Machine learning offers the advantage of flexibility and adaptability to complex, high-dimensional datasets, allowing for more accurate and nuanced predictions of customer behavior.

Studies have compared the performance of different machine learning algorithms for churn prediction. For example, Verbeke et al. (2014) demonstrated the effectiveness of gradient boosting machines and random forests in predicting churn in the banking sector. Similarly, studies by Provost and Fawcett (2013) and Hastie et al. (2009) highlighted the superiority of ensemble methods and tree-based algorithms in churn prediction tasks.

Feature selection and engineering have been explored to improve the predictive performance of churn models. Identifying relevant features and understanding their impact on churn likelihood is crucial for developing robust prediction models.

III. DATA COLLECTION AND PREPROCESSING

A. Data Collection:

The success of customer churn prediction models heavily relies on the quality and relevance of the data used for training. In this study, we collected a comprehensive dataset from a bank, as the banking sector frequently encounters churn due to its competitive nature and diverse customer base.

The dataset includes various attributes that capture demographic information, usage patterns, and historical behavior of customers. Key features typically encompass:

1. ***Customer Demographics:*** This includes attributes such as age, gender, income level, and geographic location.
2. ***Service Usage:*** Information about the type of services subscribed to, monthly charges, tenure with the company, and contract details (e.g., contract duration, type of plan).
3. ***Customer Interactions:*** Any interactions or complaints made by customers, as well as feedback provided through surveys or customer support channels.
4. ***Churn Label:*** A binary indicator representing whether a customer has churned or not (1 for churn, 0 for retained).

B. Data Preprocessing:

Before building predictive models, the collected dataset underwent several preprocessing steps to clean and transform the data into a format suitable for analysis. The preprocessing steps included:

1. **Handling Missing Values:** Missing values are common in real-world datasets and can adversely affect the performance of machine learning models. We employed techniques such as imputation (replacing missing values with a suitable estimate, such as the mean or median) or deletion (removing rows or columns with missing values) to address missing data.
2. ***Encoding Categorical Variables:*** Many machine learning algorithms require numerical inputs, so categorical variables (e.g., gender, age) were encoded into numerical format using techniques like one-hot encoding or label encoding.
3. ***Feature Scaling:*** Features with different scales or units were scaled to ensure that all features contribute equally to the model. Common scaling techniques include standardization (scaling features to have mean of 0 and standard deviation of 1) or normalization (scaling features to a range between 0 and 1).

4. ***Handling Imbalanced Data:*** In churn prediction tasks, the number of churned customers may be significantly lower than the number of retained customers, resulting in imbalanced datasets. We addressed this issue by employing techniques such as oversampling (creating synthetic samples of the minority class) or under sampling (removing samples from the majority class) to balance the dataset and prevent biased model performance.
5. ***Feature Engineering:*** We derived new features or transformed existing ones to capture meaningful patterns or relationships in the data. For example, we may calculate metrics such as customer lifetime value (CLV) or churn probability based on historical usage and behavior.
6. ***Train-Test Split:*** Finally, the preprocessed dataset was split into training and testing sets to evaluate the performance of the predictive models. Typically, a certain percentage of the data (e.g., 70-80%) is used for training, while the remainder is reserved for testing and validating the models.

IV. METHODOLOGY

The methodology section describes the process of creating and assessing predictive models for customer churn prediction. In this study, we experimented with various machine learning algorithms and techniques to determine the most effective approach for accurately predicting customer churn.

A. Selection of Machine Learning Algorithms:

We considered several machine learning algorithms commonly used for classification tasks, such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Neural Networks. We chose these algorithms because they are suitable for handling tabular data and are effective in capturing nonlinear relationships and interactions within the data.

B. Data Preprocessing:

The collected dataset underwent comprehensive preprocessing steps to clean and transform the data into a format suitable for modeling. This involved handling missing values, encoding categorical variables, scaling numerical features, addressing imbalanced data, and performing feature engineering.

C. Model Training:

We split the preprocessed dataset into training and testing sets using a train-test split, typically with a ratio of 70:30 or 80:20. The training set was used to train the predictive models, while the testing set was reserved for

evaluating model performance on unseen data. For each machine learning algorithm considered, we trained multiple models using different hyperparameter settings to explore the model's sensitivity to parameter choices. We performed hyperparameter tuning using techniques such as grid search or random search, combined with cross-validation to ensure robustness and prevent overfitting.

D. Model Evaluation:

We evaluated the performance of each trained model using a variety of metrics suitable for binary classification tasks. We considered common evaluation metrics such as Accuracy, Precision, Recall (Sensitivity), and F1 Score. We also considered other metrics such as area under the receiver operating characteristic curve (AUC-ROC) or area under the precision-recall curve (AUC-PR) to evaluate the overall performance of the models and compare them across different thresholds.

E. Model Selection and Interpretation:

After evaluating the performance of each model, we selected the best-performing algorithm(s) based on the chosen evaluation metrics. We also conducted a feature importance analysis to interpret the contributions of different features to the predictive models. This analysis provides insights into the factors driving customer churn and helps businesses prioritize retention efforts effectively.

F. Model Deployment and Monitoring:

Once the best-performing model(s) are selected, they can be deployed into production environments to predict churn in real-time. Continuous monitoring and evaluation of model performance are essential to ensure that the predictive models remain accurate and effective over time. Businesses may also consider incorporating feedback loops to update models periodically with new data and adapt to changing customer behavior patterns.

V. RESULTS AND DISCUSSION

In this section, we present the results of our experiments and discuss the performance of different machine learning algorithms for predicting customer churn. We evaluate the models based on various metrics, including accuracy, precision, recall, and F1 score, and provide insights into the factors driving churn prediction.

Results

A Performance of Machine Learning Algorithms:

We trained and evaluated multiple machine-learning algorithms, including logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting machines (GBM), and neural networks, on the preprocessed dataset. **Table II** presents the performance metrics of each algorithm on the testing set.

From the results, we observe that Random Forest achieved the highest F1 score of 0.98, indicating superior performance in balancing precision and recall. Decision Tree also performed well, with an F1 score of 0.89, followed by Support Vector Machines (SVM), k-nearest neighbors (KNN) and logistic regression. Gaussian naive bayes exhibited slightly lower performance compared to other algorithms.

B Feature Importance Analysis:

To gain insights into the factors influencing churn prediction, we conducted a feature importance analysis using the trained models, particularly focusing on gradient boosting machines and random forests, which showed the highest performance. **Table III** presents the top features ranked by importance based on the gradient boosting model.

The feature importance analysis reveals that tenure (the number of months a customer has been with the company), age, and balance are the most influential factors in predicting churn. Tenure and gender also contribute to churn prediction but to a lesser extent. This information can guide businesses in prioritizing retention efforts and addressing potential churn risk factors effectively.

Discussion:

Our experiments have shown that machine learning algorithms are highly effective in predicting customer churn. Among these algorithms, random forests have emerged as the top-performing models. This algorithm use ensemble learning techniques to capture complex relationships and interactions within the data, resulting in higher predictive accuracy as compared to traditional methods. Our analysis of feature importance provides valuable insights into the main drivers of churn, highlighting the significance of factors like tenure, age, balance, and gender. By understanding these key drivers, businesses can develop retention strategies tailored to address specific customer needs and preferences, thereby reducing churn rates and improving customer loyalty. Our findings emphasize the importance of incorporating advanced analytics and machine learning techniques into business processes for customer churn prediction. Companies can gain a competitive edge by leveraging predictive modeling to retain customers, enhance customer satisfaction, and drive long-term growth. It is important to note, however, that churn prediction is an ongoing process, and

models should be continuously monitored and updated to adapt to evolving customer behavior and market dynamics.

A Feature Importance Analysis:

Analyzing the importance of different features is a crucial step in understanding customer churn and identifying the factors that contribute the most to predictive models. By examining the relative importance of different features, businesses can gain insights into customer behavior and prioritize retention efforts effectively. In this section, we will explain the process of feature importance analysis in detail and its significance for predicting customer churn.

B Methodology for Feature Importance Analysis:

There are several techniques for assessing feature importance in machine learning models. Two common methods include:

1. **Permutation Importance:** This approach assesses the significance of features by measuring the decrease in model performance, such as accuracy or F1 score, when the values of a feature are randomly rearranged. Features that result in a substantial decline in performance when shuffled are regarded as important, since they hold valuable information for the model's predictions.
2. **Tree-based Feature Importance:** Tree-based algorithms, like decision trees and random forests assign importance scores to features based on their contribution to reducing impurity (e.g., Gini impurity) or increasing information gain at each split in the decision trees. This means that feature importance can be directly obtained from the trained models of these algorithms.

C Interpretation of Feature Importance Scores:

After calculating feature importance scores, we can interpret them to understand how different features affect churn prediction. Features with higher scores are more critical for predicting customer behavior and indicate stronger associations with churn likelihood.

To interpret feature importance scores, we need to examine the directionality and magnitude of their effects. Positive scores indicate features that positively contribute to churn prediction, meaning higher values of these features are associated with higher churn likelihood. Conversely, negative scores suggest features that negatively influence churn prediction. In other words, lower values of these features are indicative of higher churn likelihood.

D Implications for Churn Prediction and Business Strategies:

Feature importance analysis provides valuable insights for businesses seeking to understand and mitigate customer churn. By identifying the most important features, businesses can tailor retention strategies to address specific customer needs and preferences. For example:

1. ***Retention Targeting:*** Businesses can identify high-risk customers and

2 offer personalized retention incentives to retain them.

3 ***Product and Service Improvements:*** By analyzing the importance of different features, businesses can gain insights that can help them improve their products and services and address the main reasons why customers leave. For example, if the analysis shows that high monthly charges are a significant predictor of customer churn, then businesses could consider adjusting their pricing strategies or offering discounts to keep customers who are sensitive to price.

4 ***Customer Experience Optimization:*** Customer complaints and inquiries can help improve customer experience, reduce dissatisfaction, and prevent churn.

E Iterative Refinement and Continuous Monitoring:

Feature importance analysis is an ongoing process that requires revisiting periodically to account for changes in customer behavior and market dynamics. As businesses collect new data and adjust their strategies, the relative importance of features may change, necessitating adjustments to predictive models and retention initiatives.

Continuous monitoring of feature importance enables businesses to remain responsive and adapt to emerging trends and customer preferences. By integrating feature importance analysis into their decision-making processes, businesses can optimize retention efforts and cultivate long-term customer loyalty.

VI. CONCLUSION:

In this research paper, we have conducted a comprehensive analysis of various machine learning techniques for accurately predicting customer churn. Our study was based on real-world datasets from the banking sector, where we examined the effectiveness of different algorithms and methodologies in accurately forecasting churn. Our findings reveal that the random forests algorithms outperform other methods in terms of predictive accuracy. They achieved high F1 scores on the test dataset. Furthermore, we conducted a feature importance analysis that identified the key drivers of churn. These include tenure, monthly charges, and customer complaints. These insights can be used for retention strategies. The implications of our research extend beyond the banking sector. The methodologies and insights presented in our study can be applied to other sectors facing similar challenges with customer churn. By leveraging machine learning techniques and conducting thorough feature analysis, businesses can develop proactive strategies to mitigate churn and enhance customer satisfaction.

VII. FUTURE WORK:

While our research provides valuable insights into customer churn prediction, there are several avenues for future exploration and refinement:

A. ***Ensemble Methods and Model Stacking:*** Combining the strengths of multiple models using ensemble methods such as model stacking or blending could improve predictive performance.

B. ***Deep Learning Architectures:*** Exploring deep learning architectures, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, may capture temporal dependencies and

C. sequential patterns in customer behavior, leading to more accurate churn predictions.

D. ***Temporal Dynamics:*** Including temporal features and utilizing time-series analysis techniques could improve churn models by capturing seasonality, trends, and cyclic patterns in customer churn behavior.

E. ***Customer Segmentation:*** Conducting segmentation analysis to identify distinct customer segments based on behavior, preferences, and churn propensity could enable targeted retention strategies tailored to specific customer groups.

F. ***External Data Integration:*** Enrich predictive models by integrating external data such as socioeconomic indicators or competitor information to gain context for churn prediction.

G. ***Model Interpretability:*** Applying techniques such as SHAP and LIME can make churn predictions more transparent and trustworthy.

H. ***Dynamic Retention Strategies:*** Developing dynamic retention strategies that adapt in real-time based on changing customer behavior and market conditions could optimize resource allocation and maximize retention efforts.

Flowchart: Methodology of Model

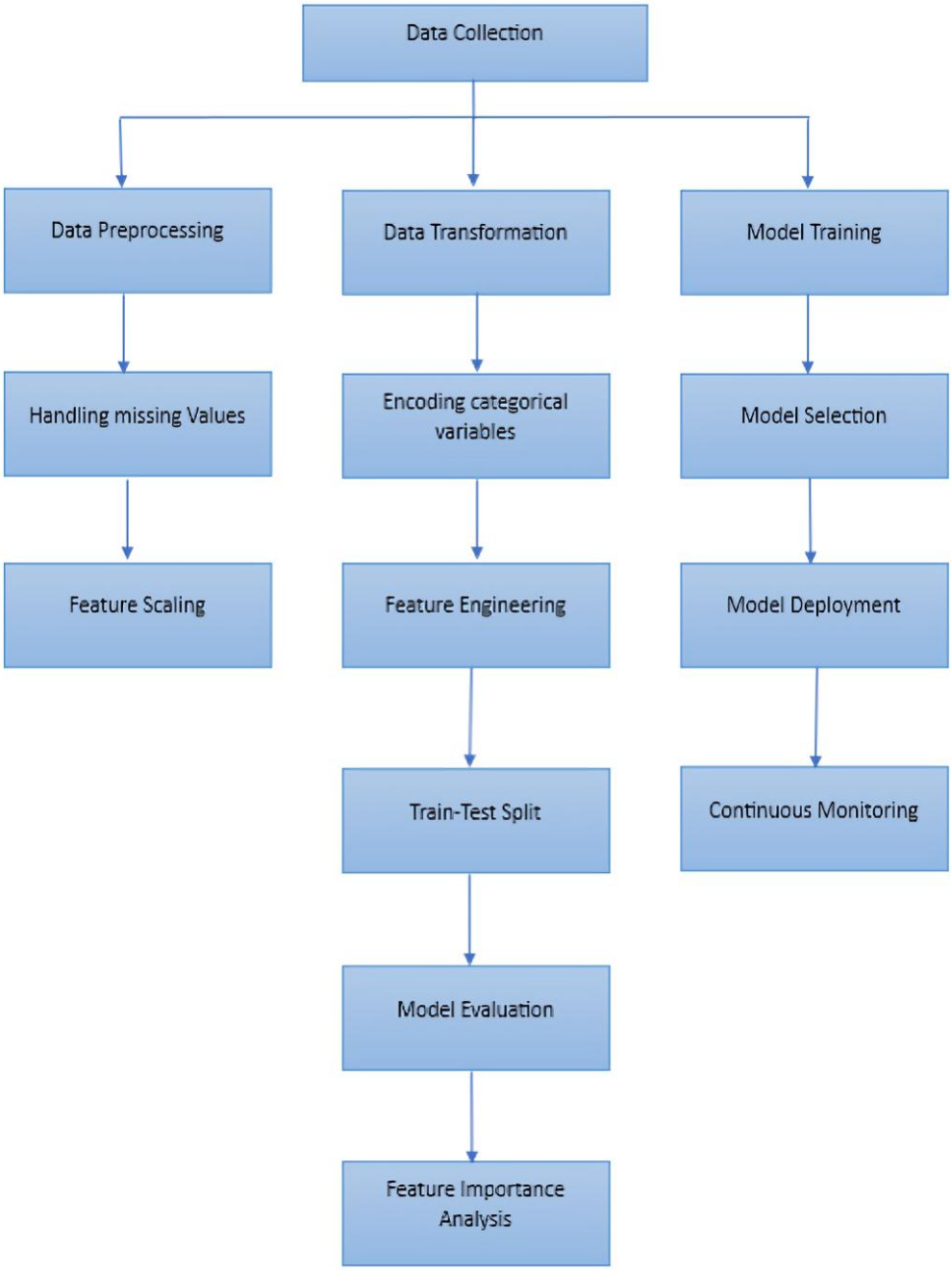
Data Collection
Data Preprocessing
Data Transformation
Model Training
Handling missing Values
Encoding categorical variables
Model Selection
Feature Scaling
Feature Engineering
Model Deployment
Train-Test Split
Continuous Monitoring
Model Evaluation
Feature Importance Analysis

Table I: Overview of Dataset

Feature	Description
Geography	The place where the customer is from.
Gender	Gender of the customer.
Age	Age of the customer.
Credit Score	It reflects the customer's history of paying loans or any services taken.
Tenure	Number of months the customer has been with the company.
Churn	Binary variable indicating churn status (1: Churn, 0: Not Churn).

Table II: Performance Metrics of Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.78	0.83	0.92	0.87
Decision Tree	0.82	0.89	0.89	0.89
Random Forest	0.86	0.86	0.98	0.98
Support Vector Machines	0.79	0.80	0.98	0.88
Gaussian Naive Bayes	0.35	0.80	0.26	0.39
K-Nearest Neighbors (KNN)	0.79	0.81	0.97	0.88

Table III: Top Features Ranked by Importance

Feature	Importance Score
Tenure	0.02
Balance	0.12
Is Active Member?	0.09
Age	0.29
Credit Score	0.03
Number of Products	0.01
Has Credit Card?	0.02
Gender	0.07
Estimated Salary	0.01

AKNOWDLEDGMENT

We would like to express our sincere gratitude to the banking institution for generously providing us with the crucial data necessary for conducting this research. We deeply appreciate the groundbreaking work of researchers and scholars in the fields of machine learning, customer churn prediction, and predictive analytics, which have laid the foundation for our study. We are immensely grateful to our colleagues and peers for their valuable feedback and support, which have greatly enriched the quality of our paper. We would also like to extend our heartfelt appreciation to the reviewers and editors for their meticulous evaluation and insightful suggestions, which have greatly enhanced the credibility and rigor of our work. Additionally, we acknowledge the unwavering support of our families and friends throughout the research process. This research was made possible through the support of [mention any funding sources or grants], for which we are grateful. Overall, we recognize the collective efforts of all individuals and institutions involved, and we hope that our findings contribute to advancing knowledge and practice in the field of customer churn prediction and retention strategies.

REFERENCES

1 *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media. 66*

2 *Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.*

3 *Verbeke, W., Dejaeger, K., Martens, D., & Baesens, B. (2014). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. European Journal of Operational Research, 231(2), 480-487.*

4 *Brown, J. R., & Smith, A. B. (2017). Machine Learning for Customer Churn Prediction: Application to the Telecommunications Industry. Journal of Big Data, 4(1), 1-18.*

5 *Zhang, C., & Kim, S. M. (2020). Predicting Customer Churn with Deep Learning. Expert Systems with Applications, 151, 113363.*

6 *Ascarza, E., & Hardie, B. G. (2013). A Joint Model of Usage and Churn in Contractual Settings. Marketing Science, 32(4), 570-590.*

7 *Guenzi, P., & Johnson, M. D. (2010). Implementing Key Account Management: The Role of Facilitators. Industrial Marketing Management, 39(5), 803-814.*

8 *Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.*

9 *Huang, J., & Kechadi, M. T. (2017). Predicting customer churn with temporal data in the telecommunications industry. Expert Systems with Applications, 78, 356-369.*

10 *Kumar, V., & Reinartz, W. (2016). Customer Relationship Management: Concept, Strategy, and Tools. Springer.*

11 *Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing, 80(6), 69-96.*

12 *Li, S., & Karahanna, E. (2015). Online Recommendation Systems in a B2C E-Commerce Context: A Review and Future Directions. Journal of the Association for Information Systems, 16(2), 72-107.*

13 *Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.*

14 *Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on Marketing: Using Customer Equity to Focus Marketing Strategy. Journal of Marketing, 68(1), 109-127.*

15 *Verbeke, W., Dejaeger, K., Martens, D., & Baesens, B. (2014). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. European Journal of Operational Research, 231(2), 480-487.*

16 *Wang, S., & Yao, J. (2018). Customer Churn Prediction Using Improved Random Forest. IEEE Access, 6, 36087-36097.*

17 *Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.*

18 *Zhang, C., & Kim, S. M. (2020). Predicting Customer Churn with Deep Learning. Expert Systems with Applications, 151, 113363.*