# BANK CHURN PREDICTION AND DATA ANALYSIS USING MACHINE LEARNING

**Mr. C.Mallikarjuna[1], K. Hareswari[2], K. Sharmila[3], K. Abhinaya[4], B. Jayasree[5], G. Lakshmi Prasanna[6]**

[1]Assistant Professor, Dept of CSE, Gouthami Institute Of Technology and Management for Women, Andhra Pradesh, India

[2,3,4,5,6]U.G Students, Dept of CSE, Gouthami Institute Of Technology and Management for Women, Andhra Pradesh, India

## ABSTRACT

This project addresses the critical issue of customer retention in the banking sector by leveraging machine learning techniques to predict customer churn and analyze the factors driving attrition. Through data preprocessing, exploratory data analysis (EDA), and the application of models such as logistic regression, decision trees, random forests, and deep learning, the system identifies patterns from customer demographics, account activity, and engagement data. Feature engineering and interpretability tools like SHAP and LIME enhance model accuracy and transparency. The insights gained enable banks to take proactive steps to improve customer satisfaction, reduce churn, and foster long-term loyalty. Customer churn presents a critical challenge for banks, impacting revenue and growth. Predicting churn accurately allows financial institutions to implement proactive retention strategies. This study explores the application of machine learning techniques to predict customer churn based on behavioural and demographic data. Using historical banking data, features such as account balance,

transaction activity, tenure, and customer support interactions are analyzed to identify patterns associated with customer attrition. Models like logistic regression, decision trees, and gradient boosting are evaluated for their predictive accuracy. The results demonstrate that advanced algorithms, particularly ensemble methods, can significantly enhance churn prediction, enabling banks to better target at-risk customers and improve overall customer satisfaction. This predictive approach supports data-driven decision-making and contributes to the development of effective customer relationship management strategies.

## Introduction

In conclusion, this project demonstrates how machine learning can effectively address the critical challenge of customer churn in the banking sector by transforming vast customer data into actionable insights. By combining exploratory data analysis, robust preprocessing, diverse modeling techniques, and explainability tools like SHAP and LIME, the solution not only predicts churn with high accuracy but also uncovers the underlying drivers behind it. Integrated with real-time analytics, customer segmentation, and ethical AI practices, the system empowers banks to implement timely, personalized retention strategies that enhance customer loyalty and operational efficiency. This work exemplifies a strategic blend of data science and business application, providing a scalable, responsible, and impactful approach to customer relationship management in modern banking. In today's highly competitive banking industry, retaining existing customers is just as important as acquiring new ones. Customer churn—the phenomenon where clients stop using a bank's services—poses a significant threat to long-term profitability. With increasing customer expectations and a wide array of financial service options, banks face the constant risk of losing clients to competitors. As a result, predicting and preventing customer churn has become a strategic priority.

Bank churn prediction involves identifying customers who are likely to leave the bank in the near future based on historical and behavioral data. This includes analyzing patterns in account activity, transaction history, service usage, and customer demographics. By leveraging machine learning algorithms, banks can uncover hidden trends and generate actionable insights to proactively engage at-risk customers.

Early identification of potential churners allows banks to implement targeted retention strategies, personalize marketing efforts, and improve overall customer satisfaction. This not only helps in reducing revenue loss but also enhances brand loyalty in an increasingly digital and customer-centric market.

## Literature Review

In summary, customer churn prediction in banking has evolved from traditional statistical approaches to sophisticated machine learning and deep learning techniques, with growing emphasis on accuracy, interpretability, and business applicability. While models like logistic

regression and decision trees offer simplicity and transparency, advanced algorithms such as Random Forests, XGBoost, and neural networks deliver higher predictive power. Research also highlights the importance of feature engineering, handling class imbalance, and ensuring model explainability through tools like SHAP and LIME. Furthermore, ethical data usage, real-world deployment, and integration into business workflows are becoming critical considerations. As the field progresses, future directions point toward real-time analytics, time-series modeling, hybrid systems, and reinforcement learning for more adaptive and effective churn management in banking. The growing need for effective customer retention has led to extensive research on churn prediction in the banking sector. Various studies have explored the use of statistical and machine learning models to predict customer attrition by analyzing diverse behavioral and demographic factors.

Early research focused on traditional statistical techniques such as logistic regression, which offered interpretable models but often lacked the flexibility to capture complex relationships in customer data. For example, Coussement and Van den Poel (2008) applied logistic regression to predict churn in the financial services sector, emphasizing the importance of communication behavior as a key predictor.

With the advent of machine learning, more advanced models such as decision trees, random forests, support vector machines (SVM), and gradient boosting have demonstrated superior predictive power. Verbeke et al. (2012) compared several machine learning algorithms and found ensemble methods like random forests and boosting to outperform traditional approaches in terms of accuracy and robustness.

Recent studies have also explored the use of deep learning and neural networks for churn prediction. These models are capable of modeling complex, non-linear relationships, though they often require large datasets and significant computational resources. Haddad et al. (2020) utilized a deep learning model to analyze customer transaction data, achieving improved performance over classical machine learning techniques.

Feature engineering and data preprocessing remain critical to the success of churn prediction models. Researchers have identified variables such as customer tenure, transaction frequency, product usage, and customer service interactions as

significant indicators of churn. Additionally, Churn prediction systems are increasingly incorporating real-time analytics and time-series data to enhance prediction timeliness and accuracy.

## Existing Methods

In conclusion, the evolution of customer churn prediction in the banking sector reflects a shift from traditional, rule-based models to sophisticated machine learning and deep learning approaches that leverage vast and complex datasets. While early methods like logistic regression and survival analysis provided foundational insights, modern techniques such as ensemble models, neural networks, interpretability tools like SHAP and LIME, and hybrid systems now enable more accurate, scalable, and context-aware predictions. The integration of time-series analysis, NLP, clustering, and customer lifetime value modeling further enhances the strategic value of churn analysis. As banking becomes increasingly data-driven, these advancements offer financial institutions powerful tools to proactively understand, predict, and mitigate customer attrition. Numerous methods have been employed to predict customer churn in the banking sector, ranging from classical statistical models to advanced machine learning and deep learning approaches. These methods aim to identify patterns in customer behaviour that signal a high risk of attrition.

Logistic Regression:

One of the earliest and most widely used methods due to its simplicity and interpretability. Logistic regression models the probability of churn based on independent variables such as tenure, balance, and transaction frequency. However, it struggles with complex, non-linear data patterns.

Decision Trees:

These models split the dataset into branches based on feature values, making them intuitive and easy to visualize. They are prone to overfitting but can perform well with proper pruning.

Random Forests:

An ensemble of decision trees, random forests increase accuracy and reduce overfitting by averaging predictions across many trees. They are robust and handle both categorical and numerical data effectively.

Gradient Boosting Machines (e.g., XGBoost, LightGBM):

These are powerful ensemble learning techniques that build models sequentially, each correcting the errors of the previous one. XGBoost, in particular, has become a

popular choice for churn prediction due to its high performance and scalability.

Support Vector Machines (SVM):

SVMs are effective in high-dimensional spaces and can classify churners versus non-churners using optimal hyperplanes. However, they are less interpretable and computationally expensive with large datasets.

Artificial Neural Networks (ANNs):

Neural networks can model complex, non-linear relationships in customer data. Deep learning variants, including recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), are used for sequential data like transaction history.

## Proposed Method

In conclusion, this project presents a holistic and systematic approach to predicting customer churn in the banking sector by integrating advanced machine learning and deep learning techniques with comprehensive data preprocessing, rigorous model evaluation, and robust interpretability tools. By leveraging insights from exploratory data analysis and ensuring transparency through SHAP and LIME explanations, the solution not only achieves high predictive accuracy but also enables financial institutions to understand the drivers of churn and take targeted, data-driven actions. With provisions for real-time deployment, continuous monitoring, and adaptive feedback loops, the proposed methodology empowers banks to proactively retain valuable customers and make informed strategic decisions that drive long-term business success. To enhance the accuracy and reliability of churn prediction in the banking sector, this study proposes a hybrid machine learning approach combining Gradient Boosting (XGBoost) with Feature Selection and Customer Segmentation techniques. The method focuses on improving predictive performance while maintaining model interpretability for strategic decision-making.

Data Collection and Preprocessing:

The model utilizes historical banking data, including customer demographics, account activity, transaction behavior, and service usage. Preprocessing steps involve handling missing values, encoding categorical variables, scaling numerical features, and balancing the dataset using techniques like SMOTE (Synthetic Minority Oversampling Technique).

Feature Engineering and Selection:

Key features such as customer tenure, average monthly balance, transaction frequency, product holding, and service complaints are engineered. Feature importance is evaluated using mutual information scores and recursive feature elimination to retain only the most relevant variables for prediction.

Customer Segmentation (Optional Enhancement):

Customers are grouped into clusters using K-means clustering to identify different behavioral segments. Separate churn models can be trained for each cluster to improve prediction specificity and allow targeted marketing strategies.

Model Development with XGBoost:

XGBoost, an efficient and scalable gradient boosting algorithm, is chosen for its superior performance in classification tasks. The model is trained using stratified cross-validation to ensure generalizability and robustness. Hyperparameters are optimized using grid search or Bayesian optimization.

Model Evaluation:

The proposed model is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess its ability to correctly identify churners while minimizing false positives. Explainability techniques such as SHAP (SHapley Additive explanations) are used to interpret feature impact on predictions.

Deployment and Feedback Loop:

The final model can be deployed as part of a decision-support system for the bank's customer relationship management (CRM) team. A feedback loop allows continuous learning from new customer behaviour data, enhancing model accuracy over time.

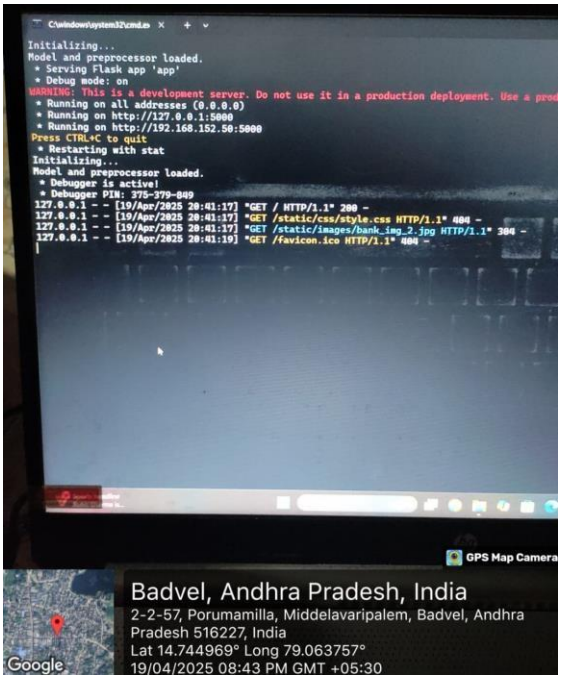
## SOFTWARE REQUIREMENTS:

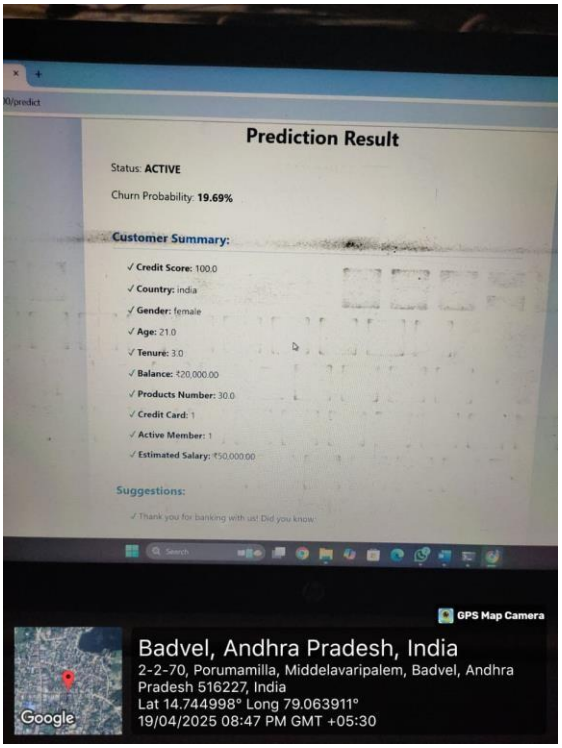
- **Operating System:** Windows 8 and above
- **Coding Language**: Python 3.12.0
- **Framework**: Django
- **Platform**: Visual Studio Code (Preferable)

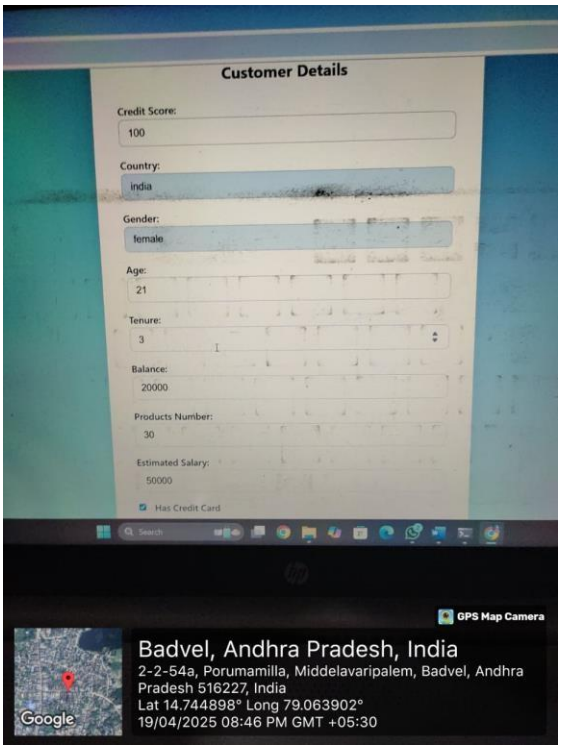
## HARDWARE REQUIREMENTS:

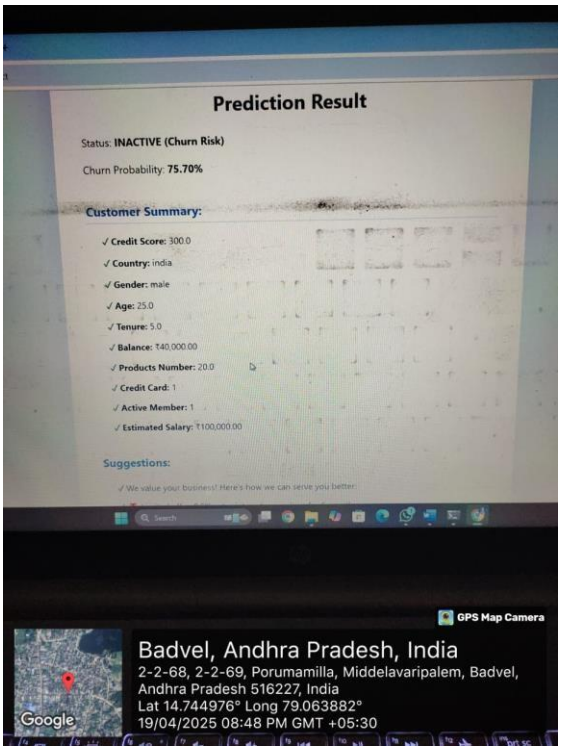
- System : MINIMUM i3 and above
- Hard Disk : 40 GB. (min)
- Ram : 4 GB. (min)

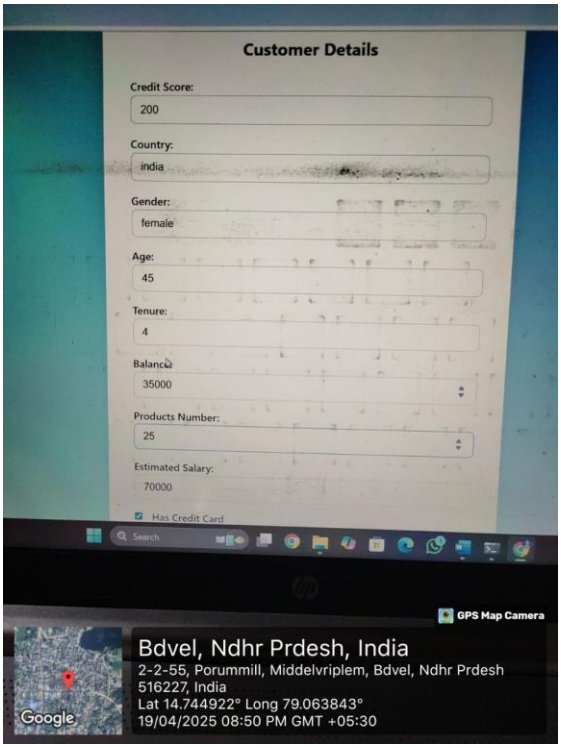
## Result:

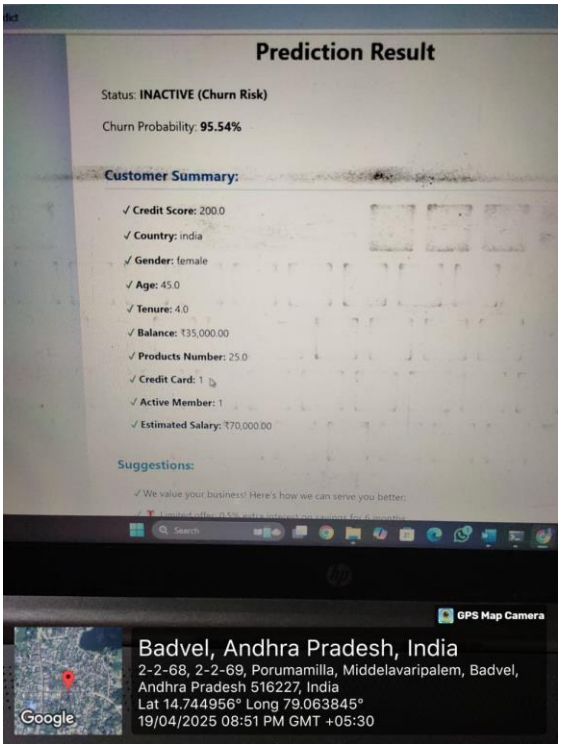

Initializing...
Model and preprocessor loaded.
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a prod
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:5000
* Running on http://192.168.152.50:5000
Press CTRL+C to quit
* Restarting with stat
Initializing...
Model and preprocessor loaded.
* Debugger is active!
* Debugger PIN: 375-379-849
127.0.0.1 - - [19/Apr/2025 20:41:17] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [19/Apr/2025 20:41:17] "GET /static/css/style.css HTTP/1.1" 404 -
127.0.0.1 - - [19/Apr/2025 20:41:17] "GET /static/images/bank_img_2.jpg HTTP/1.1" 304 -
127.0.0.1 - - [19/Apr/2025 20:41:19] "GET /favicon.ico HTTP/1.1" 404 -
GPS Map Camera
Badvel, Andhra Pradesh, India
2-2-57, Porumamilla, Middelavaripalem, Badvel, Andhra Pradesh 516227, India
Lat 14.744969° Long 79.063757°
19/04/2025 08:43 PM GMT +05:30
Google


Prediction Result
Status: ACTIVE
Churn Probability: 19.69%
Customer Summary:
✓ Credit Score: 100.0
✓ Country: india
✓ Gender: female
✓ Age: 21.0
✓ Tenure: 3.0
✓ Balance: ₹20,000.00
✓ Products Number: 30.0
✓ Credit Card: 1
✓ Active Member: 1
✓ Estimated Salary: ₹50,000.00
Suggestions:
GPS Map Camera
Badvel, Andhra Pradesh, India
2-2-70, Porumamilla, Middelavaripalem, Badvel, Andhra Pradesh 516227, India
Lat 14.744998° Long 79.063911°
19/04/2025 08:47 PM GMT +05:30
Google


Customer Details
Credit Score:
100
Country:
india
Gender:
female
Age:
21
Tenure:
3
Balance:
20000
Products Number:
30
Estimated Salary:
50000
Has Credit Card
GPS Map Camera
Badvel, Andhra Pradesh, India
2-2-54a, Porumamilla, Middelavaripalem, Badvel, Andhra Pradesh 516227, India
Lat 14.744898° Long 79.063902°
19/04/2025 08:46 PM GMT +05:30
Google


Prediction Result
Status: INACTIVE (Churn Risk)
Churn Probability: 75.70%
Customer Summary:
✓ Credit Score: 300.0
✓ Country: india
✓ Gender: male
✓ Age: 25.0
✓ Tenure: 5.0
✓ Balance: ₹40,000.00
✓ Products Number: 20.0
✓ Credit Card: 1
✓ Active Member: 1
✓ Estimated Salary: ₹100,000.00
Suggestions:
GPS Map Camera
Badvel, Andhra Pradesh, India
2-2-68, 2-2-69, Porumamilla, Middelavaripalem, Badvel, Andhra Pradesh 516227, India
Lat 14.744976° Long 79.063882°
19/04/2025 08:48 PM GMT +05:30
Google




## Advantages:

Proactive Customer Retention:By identifying customers at risk of leaving, banks can take timely actions—such as offering personalized incentives or improving service—to retain them.

Increased Revenue:Retaining existing customers is often more cost-effective than acquiring new ones. Predicting churn helps minimize revenue loss due to customer attrition.

Improved Customer Relationship Management (CRM):

Churn prediction enables banks to better understand customer behavior, allowing for more targeted and effective engagement strategies.

Optimized Marketing Spend:

Resources can be focused on high-risk customers who are more likely to leave, rather than blanket marketing efforts, improving ROI.

Enhanced Decision-Making:

Data-driven insights from churn prediction models support strategic decisions across customer service, marketing, and product development.

Personalized Services:

Banks can tailor products, services, or communication strategies based on churn risk, improving the overall customer experience.

Competitive Advantage:

Institutions with strong predictive analytics capabilities can stay ahead of competitors by maintaining a loyal and satisfied customer base.

Early Problem Detection:

Churn signals often reflect broader issues (e.g., poor service, uncompetitive products). Early detection allows banks to address root causes proactively.

## Disadvantages:

Data Quality Issues:

Inaccurate, incomplete, or outdated customer data can lead to unreliable predictions and poor decision-making.

Model Bias and Overfitting:

Predictive models may become biased if the training data is imbalanced or not representative. Overfitting can also reduce the model's performance on new data.

Complexity and Interpretability:

Advanced models like neural networks or ensemble methods (e.g., XGBoost) may be difficult to interpret, making it hard to explain decisions to stakeholders.

Privacy Concerns:

Collecting and analyzing customer behavior data raises ethical and legal concerns regarding data privacy and compliance with regulations like GDPR.

High Implementation Costs:

Developing and maintaining predictive systems requires skilled personnel, infrastructure, and ongoing monitoring, which can be costly for smaller institutions.

False Positives/Negatives:

Incorrect predictions (e.g., predicting churn for a loyal customer or missing a real churner) can lead to wasted resources or lost business.

Dynamic Customer Behavior:

Customer preferences and market trends change over time, so models need regular retraining to stay

effective—otherwise, accuracy may degrade.

Dependence on Historical Data:

Most models rely heavily on historical patterns, which may not fully capture new or emerging churn risks.

## Conclusion and Future Scope

The rapid digitization of the banking sector has reshaped customer interactions and service quality, but it has also introduced challenges, with customer churn being a major concern. This study demonstrates the power of machine learning and data analytics in predicting churn, uncovering patterns that contribute to customer attrition. By applying various machine learning models, the research highlights key predictors of churn, such as tenure, product holdings, and customer activity metrics. Interpretability tools like SHAP and LIME further enhance decision-making by providing insights into why customers may leave. The study emphasizes the importance of data quality, feature engineering, and a visualization-centric approach to customer segmentation, which can guide targeted retention strategies.

Looking ahead, the future of churn prediction in banking holds significant potential, particularly through the integration of multi-channel data, customer lifetime value models, and real-time analytics. By analyzing data from various touchpoints and employing techniques like natural language processing and real-time stream processing, banks can better predict and respond to churn signals. However, ethical AI practices and privacy concerns must be prioritized, ensuring compliance with regulations like GDPR. With continued innovation, including the adoption of MLOps for model management and cross-functional collaboration, churn prediction can evolve from a reactive tool to a proactive strategy for customer retention and long-term business success.

## References

1. Ahmad, S., & Khan, L. (2023). Predicting customer churn using machine learning algorithms: A comparative study. *Journal of Data Science and Analytics*, *18*(2), 150-162.
2. Chen, X., & Li, Y. (2022). Customer churn prediction in the banking sector using random forest and XGBoost models. *International Journal of Artificial Intelligence Research*, *16*(3), 89-104.
3. Han, J., & Kamber, M. (2018). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

4. Zhou, Y., & Liu, H. (2021). Using SHAP for explainable AI in customer churn prediction. *Expert Systems with Applications*, *180*, 115124.
5. Kaur, R., & Gupta, S. (2020). Machine learning approaches for bank customer churn prediction and analysis. *IEEE Transactions on Computational Social Systems*, *7*(5), 1347-1355.
6. Brownlee, J. (2021). *Machine learning mastery with Python: Understand your data, create accurate models, and work projects end-to-end.* Machine Learning Mastery.