Analysing the financial future of Banks

BHARATH CHANDRA, SRI PAVANI MANIKONDA, PESALA NAGA SASHANK, MEETALI KEWLAN, NAMYA ARORA, PIYUSH RANA, RICHIK MOHAN

1. Abstract:

In the banking sector, customer churn is a pressing challenge with significant implications. This project utilizes supervised learning algorithms to create a predictive model aimed at identifying at-risk customers and uncovering the reasons behind their attrition. Through a thorough dataset exploration, it reveals patterns and indicators signaling potential churn, with the goal of accurately forecasting customer attrition. The project has dual benefits: proactive identification of high-risk customers for targeted retention efforts and systematic resolution of churn drivers, enhancing customer satisfaction and loyalty. It exemplifies the practical application of data-driven decision-making in banking, offering a blueprint for leveraging modern analytics and machine learning to address the critical issue of customer churn.

2. Keywords:

1. **Churn Prediction:** Churn prediction refers to the process of using data analysis and predictive modeling to identify customers or users who are likely to discontinue their relationship with a service or product, such as canceling a subscription or leaving a bank.

2. **Machine Learning:** Machine learning is a branch of artificial intelligence where algorithms are used to enable computers to learn from data and make predictions or decisions without being explicitly programmed, often applied in tasks like churn prediction.

3. **SVM (Support Vector Machine):** SVM is a machine learning algorithm used for classification and regression tasks. It finds a hyperplane that best separates data into different classes, making it useful in binary churn prediction.

4. **Logistic Regression:** Logistic regression is a statistical model used for binary classification tasks, like churn prediction. It models the probability of an event occurring based on input variables.

5. **Random Forest:**Random forest is an ensemble learning technique that combines multiple decision trees to make more accurate predictions. It's often employed in churn prediction due to its robustness and accuracy.

6. **Decision Tree:** A decision tree is a supervised machine learning model that maps decisions or outcomes based on input features. It is used in churn prediction to create a flowchart-like structure for decision-making.

7. **EDA (Exploratory Data Analysis):** EDA is the process of visually and statistically analyzing data sets to summarize their main characteristics, often used as a preliminary step in churn prediction to gain insights into the data's structure and potential patterns.

3.Introduction:

Bank churning is a financial strategy that involves utilizing different incentives and bonuses provided by banks to open new accounts. The term 'churning' is aptly named because individuals who practice it often open new bank accounts multiple times, thereby 'churning' through various banking institutions. This practice is entirely legal and is gaining popularity due to several advantages it offers over traditional investment strategies, while also having few, if any, downsides.

The array of benefits provided by bank churning is one of its primary attractions. Banks frequently attract customers by offering lucrative sign-up bonuses, cash rewards, high-interest rates on savings accounts, and other financial benefits. By strategically timing bank account openings and closings, churning enthusiasts can potentially earn substantial sums of money. Moreover, bank churning requires minimal financial risk compared to other investment options, which makes it an appealing option for those looking to increase their wealth without significant exposure to market volatility.

The accessibility of bank churning is another important advantage. Bank churning is accessible to a broad range of individuals, unlike complex investment strategies that require in-depth financial knowledge. By taking advantage of promotions and bonuses offered by financial institutions, individuals can maximize their daily banking activities.

Although bank churning has many advantages, it's important to take a responsible approach to it. Successful churning requires meticulous planning, tracking account activity, and adhering to bank policies and regulations. Individuals should weigh the potential benefits against their personal financial goals and risk tolerance before starting bank churning, just as they would do with any financial strategy.

4.PROBLEM STATEMENT:

The objective of this project is to develop a predictive model that is both accurate and reliable, and can predict a bank's growth trajectory over the next few years. By leveraging historical financial and operational data, the aim is to identify key indicators and patterns that can help stakeholders make informed decisions about strategic planning, resource allocation, and risk management. The predictive model's approach involves using customer churn, which assists the bank in maintaining their customers.

Logistic Regression, SVM, Decision Tree, and Random Forest classification algorithms are being used in this project, along with an explainable AI algorithm.

Dataset Description:

The dataset was extracted from the Kaggle. This data contains 12 features about 10000

clients of the bank.

The features are the following:

• customer_id, unused variable.

• credit_score, used as input.

• country, used as input.

• gender, used as input.

• age, used as input.

• tenure, used as input.

• balance, used as input.

• products_number, used as input.

• credit_card, used as input.

• active_member, used as input.

• estimated_salary, used as input.

• Excited, used as the target. 1 if the client has left the bank during some period or 0 if he/she has not. This project uses customer data from a bank to build a predictive model for the likely churn clients. As we know, it is much more expensive to sign in a new client than to keep an existing one. It is advantageous for banks to know what leads clients to leave the company.Churn prevention allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible.

5. LITERATURE SURVEY:

Literature Paper	Abstract	Pros	Cons
1. Bank efficiency and failure prediction: a nonparametric and dynamic model based on data envelopment analysis Zhiyong Li, Chen Feng, Ying Tang	This research addresses bank failure prediction by focusing on bank efficiency as a potential indicator of financial stability. Using a nonparametric method called Malmquist DEA with Worst Practice Frontier, it dynamically assesses bankruptcy risk in a dataset spanning 15 years, including the subprime financial crisis. The study compares this approach to static models and finds that Malmquist DEA is effective not only in estimating productivity growth but also in providing early warnings of potential bank collapses. Different variations of the DEA model also demonstrate strong predictive capabilities, highlighting its value in assessing and predicting bank failures over time.	Innovative Methodology: The paper introduces a novel approach, Malmquist DEA with Worst Practice Frontier, for dynamically assessing the bankruptcy risk of banks over multiple periods, offering a fresh perspective on bank failure prediction. Empirical Analysis: The study uses a substantial dataset of 4426 US banks over 15 years, including the subprime financial crisis, enhancing the credibility of its findings. Comparative Analysis: It conducts a thorough comparison with static traditional DEA models and other extended dynamic models, demonstrating the superiority of the proposed approach in predicting bank failures. Practical Relevance: The research provides valuable insights for stakeholders and regulators, especially in times of economic uncertainty like the COVID-19 pandemic.	Complexity: The methodology introduced may be relatively complex for those without a strong background in data envelopment analysis (DEA), potentially limiting its accessibility. Data Limitations: The accuracy and reliability of predictions heavily depend on the quality and completeness of the dataset, which can be a limitation in real-world applications. Generalizability: While the study focuses on US banks, the applicability of the proposed approach to banks in other countries or regions may require further validation. Interpretability: The paper could benefit from more detailed explanations and examples to enhance the understanding of the proposed methodology for a broader audience. Assumptions: Like all modeling approaches, the Malmquist DEA method makes certain assumptions that may not always hold in all banking contexts, potentially impacting the accuracy of predictions.

2. Artificial intelligence and bank credit analysis: A review Hicham Sadok Fadi Sakka Mohammed El Hadi El Maknouzi	This article teases out the ramifications of artificial intelligence (AI) use in the credit analysis process by banks and other financing institutions. The unique features of AI models, coupled with the expansion of computing power, make new sources of information (big data) available for creditworthiness assessments. Combined, the use of AI and big data can capture weak signals, whether in the form of interactions or non-linearities between explanatory variables that appear to yield prediction improvements over conventional measures of creditworthiness. At the macroeconomic level, this translates into positive estimates for economic growth. On a micro scale, instead, the use of AI in credit analysis improves financial inclusion and access to credit for traditionally underserved borrowers.	Advancement in Credit Analysis: The paper discusses the positive impact of AI and big data on credit analysis, highlighting their ability to capture subtle patterns and improve creditworthiness assessments. Macro and Micro Benefits: It explores the benefits of AI-based credit analysis at both macroeconomic and micro levels, including potential positive effects on economic growth and financial inclusion. Relevance to Banking: The article is highly relevant to the banking industry, offering insights into how AI can enhance credit risk management and customer service.	Ethical and Regulatory Concerns: The paper acknowledges enduring concerns related to AI in credit analysis, including potential biases and ethical, legal, and regulatory issues, which require careful consideration. Complexity: While the benefits of AI are discussed, the complexity of implementing AI-based credit analysis is not extensively addressed, which may pose challenges for some institutions. Generalization: The study focuses on the potential advantages of AI in a specific context, and the extent to which these findings apply to other industries or regions is not thoroughly examined.

3. THE INFLUENCE OF SUSTAINABLE INNOVATION ON FINANCIAL ENTREPRENEURSHIP PERFORMANCE: GROWTH AND PREDICTION IN AN EMERGING MARKET Madher Ebrahim Hamdallah Anan Fathi Srouji *Journal of Governance and Regulation (2022)*	This study aims to perceive the effect of financial entrepreneurship performance (FEP) over sustainable innovation (SI) disclosure in an emerging market. Jordanian banks are tested based on a multiple regression analysis for the periods 2008 and 2018 and a time series forecasting webinar analysis for the period from 2019 to 2029 based on data ranging from 2008 to 2018. Innovation is indicated through disclosed intangible assets (IA), and items related to research and development (R&D) costs. As organizations anticipate stability by concentrating on technological awareness to influence higher innovative performance (Guo, Guo, Zhou, & Wu, 2020), this study came to converse the relationships between previous literature variables; Hussain (2015) as well as Lassala, Apetrei, and Sapena (2017) revealed through the regression models that there is a relationship between FEP and SI.	Relevance: The study addresses the important relationship between financial entrepreneurship performance (FEP) and sustainable innovation (SI) disclosure, which is relevant in the context of emerging markets like Jordan. Longitudinal Analysis: The study spans over a decade (2008-2018) and includes a time series forecasting analysis for the future (2019-2029), providing a comprehensive view of how these variables evolve over time. Building on Previous Research: The study builds on the findings of previous literature (Hussain, 2015; Lassala et al., 2017), contributing to the existing body of knowledge on the relationship between FEP and SI. Quantitative Approach: The use of multiple regression analysis and time series forecasting adds robustness to the research methodology, allowing for statistical analysis and predictive modeling.	Data Limitations: The study relies on historical data from 2008 to 2018 and forecasting for future years. Data quality and availability, especially for the future, may be uncertain, potentially affecting the accuracy of predictions. Simplistic Measurement of Innovation: The study uses disclosed intangible assets (IA) and research and development (R&D) costs as indicators of innovation. These measures may not capture the full spectrum of innovation activities in banks. Narrow Focus: The study primarily focuses on the relationship between FEP and SI, leaving out potential contextual or external factors that could influence these variables. Limited Contextual Analysis: The study does not extensively discuss the specific contextual factors in Jordan that may affect FEP and SI, potentially limiting the generalizability of its findings to other regions or markets. Assumption of Causality: While the study finds a positive relationship between FEP and SI, it does not establish causality, leaving room for alternative interpretations of the relationship.

4. Cash and the Economy: Evidence from India's Demonetization Gabriel Chodorow-Reich Gita Gopinath Abhinav Narayanan	We analyze a unique episode in the history of monetary economics, the 2016 Indian "demonetization." This policy made 86% of cash in circulation illegal tender overnight, with new notes gradually introduced over the next several months. We present a model of demonetization where agents hold cash both to satisfy a cash-in-advance constraint and for tax evasion purposes. We test the predictions of the model in the cross-section of Indian districts using several novel data sets including: the geographic distribution of demonetized and new notes for causal inference; night light activity and employment surveys to measure economic activity including in the informal sector; debit/credit cards and e-wallet transactions data; and banking data on deposit and credit growth. Districts experiencing more severe	Real-World Relevance: The study examines a significant real-world event, the 2016 Indian demonetization, providing insights into its economic effects and implications for monetary policy. Unique Data: The research utilizes a diverse set of data sources, including geographic distribution of notes, night light activity, employment surveys, and banking data, enhancing the depth and breadth of the analysis. Causal Inference: The study aims to establish causality by examining how districts with varying degrees of demonetization experienced different economic outcomes, contributing to a better understanding of the policy's impact. Policy Implications: By analyzing the effects of demonetization on economic activity, adoption of alternative payment technologies, and bank credit growth, the study provides valuable insights for policymakers and economists. Empirical Testing: The research challenges the concept of monetary neutrality using a large-scale natural experiment, adding to the literature on the effects of monetary policy.	Complexity: The study involves a complex model and multiple data sources, which may make it challenging for some readers to follow or replicate the analysis. Data Limitations: While the study uses various data sets, the accuracy and representativeness of these data sources may vary, potentially affecting the robustness of the findings. Specific Context: The findings are specific to the 2016 Indian demonetization episode and may not be directly applicable to other monetary policy contexts or countries. Short-Term Focus: The analysis primarily focuses on the short-term effects of demonetization, and the longer-term consequences may not be fully addressed. Assumptions in the Model: The study's model assumes that agents hold cash for both legal and illegal purposes. The accuracy of these assumptions may influence the model's predictions and findings.

5. Artificial Intelligence and Human Psychology in Online Transaction Fraud Raheela Firdaus Yang Xue Muhammad Sibt e Ali	Banking operations have changed due to technological advancement. On one hand, modernization in technology has facilitated the daily operation of banks; on the other hand, this has also resulted in an increase in the number of cyber-attacks. Artificial Intelligence has introduced new models to detect and prevent cybercrimes. Some fraud has also occurred due to the involvement of employees inside particular organizations. So, this study has focused on both sides: the machine as well as the human. Firstly, the research has focused on fraud diamond theory and has analyzed factors such as rationalization, capabilities, perceived pressure, and perceived opportunities to understand the psychology of the fraudster. Secondly, Artificial Intelligence characteristics, threat exposure, big data management, explainability, cost effectiveness, and risk prediction are evaluated to explore their use in fraud reduction in banks.	Comprehensive Approach: The research combines both human factors (fraud diamond theory) and technological aspects (Artificial Intelligence) to provide a holistic understanding of fraud prevention and detection. Practical Insights: The study offers practical insights for the banking industry in Pakistan, providing actionable recommendations to control fraud both inside and outside organizations. Empirical Data: The research collects data from 15 banks in Pakistan, adding empirical evidence to support its findings, making it more robust and applicable to real-world scenarios. Economic Impact: By addressing fraud and cybercrimes, the study indirectly contributes to the economic growth of Pakistan by enhancing the security and stability of its banking sector	Limited Generalizability: Findings and recommendations may be specific to the context of Pakistan's banking industry and may not be directly applicable to other regions or countries with different regulatory and technological environments. Data Collection Challenges: The study relies on self-reported data collected through questionnaires, which may be subject to biases or inaccuracies, depending on respondents' understanding and willingness to disclose sensitive information. Complexity: Combining human psychology (fraud diamond theory) and technological aspects (AI) can make the study complex, potentially making it challenging for some readers to grasp the full scope of the research. Limited Discussion of Ethical Issues: The study primarily focuses on technical and psychological aspects, and there is limited discussion of the ethical considerations surrounding the use of AI for fraud detection and prevention. Future Trends: Given the rapid evolution of technology and cyber threats, the study may not fully account for future trends and challenges in the field of banking and cybersecurity.

6. *Bank capital and value in the cross-section* Hamid Mehran Anjan Thakor	We develop a dynamic model of bank capital structure in an acquisitions context which predicts: (i) total bank value and the bank's equity capital are positively correlated in the cross-section, and (ii) the various components of bank value are also positively cross-sectionally related to bank capital. Our empirical tests provide strong support for these predictions. The results are robust to a variety of alternative explanations - growth prospects, desire to acquire toe-hold positions, desire of capital-starved acquirers to buy capital-rich targets, market timing, pecking order, the effect of banks with binding capital requirements, Too Big To Fail, target profitability, risk, and mechanical effects	Empirical Support: The abstract mentions that the empirical tests provide strong support for the model's predictions. This is a significant strength as it indicates that the model is likely to be a good representation of real-world banking dynamics. Robustness: The results are said to be robust to a variety of alternative explanations. This suggests that the model is comprehensive and able to account for various factors that may influence bank capital structure in the context of acquisitions. Practical Application: Understanding the relationship between bank capital structure and acquisitions is valuable for both practitioners in the banking industry and policymakers. It can inform decision-making and regulatory frameworks.	Lack of Specifics: The abstract provides an overview of the model's predictions and its empirical support but lacks specific details about the model itself. This makes it difficult for readers to fully grasp the model's mechanics and limitations. Assumptions and Simplifications: Any model involves simplifications and assumptions about the real world. Without knowing these assumptions, it's hard to assess the model's accuracy and applicability in different scenarios. Data Limitations: The strength of empirical tests depends on the quality and quantity of data used. If the data used in the study is limited or not representative of the banking industry as a whole, it could weaken the validity of the results. Generalizability: While the abstract claims robustness to alternative explanations, it's essential to assess whether the findings of this model can be generalized to different banking contexts, regions, or time periods. Banking systems can vary widely, and what holds true in one context may not apply elsewhere. Complexity: Dynamic models of bank capital structure can be highly complex. The abstract doesn't provide any information on the model's complexity, which could make it challenging for non-experts to understand and evaluate.

7. Bank liquidity, stock market participation, and economic growth Elena Mattana Ettore Panetti *Journal of Banking and Finance (2014)*	We develop a growth model with banks and markets to reconcile the observed decreasing trend in the relative liquidity of many financial systems around the world with the increasing household participation in direct market trades. At low levels of economic development, the presence of fixed entry costs prevents the agents from accessing the market, and pushes them towards the banks, which provide high relative liquidity. We characterize the threshold after which the agents are rich enough to access the market, where the relative liquidity is lower, and show that the relative liquidity of the whole financial system (banks and markets) drops because of the increasing market participation. We provide some evidence consistent with this theoretical prediction: a one-unit increase in an index of securities market liberalization leads to a drop in the relative liquidity of between 17 and 27 per cent.	Reconciliation of Trends: The abstract addresses an interesting and important issue - the decreasing relative liquidity of financial systems alongside increasing household participation in direct market trades. This is a relevant and timely topic for financial economists and policymakers. Theoretical Model: The abstract introduces a theoretical model that attempts to explain the observed trends. Developing such models can provide valuable insights into complex economic phenomena, helping researchers and policymakers better understand the dynamics at play. Threshold Characterization: The model characterizes a threshold beyond which agents are considered "rich enough" to access the market. This concept could offer a practical framework for understanding the transition from bank-based to market-based financial systems. Empirical Evidence: The abstract mentions providing empirical evidence to support the theoretical prediction. This combination of theory and evidence is a strong point as it demonstrates an attempt to validate the model's implications.	Complexity: Economic models, especially those involving banks, markets, and liquidity, can be highly complex. The abstract doesn't provide details about the model's assumptions and mechanics, which could make it difficult for readers to assess its validity. Data and Evidence: While the abstract mentions empirical evidence, it doesn't specify the nature of the data used or the methodology employed. This lack of detail can raise questions about the quality and relevance of the evidence. Simplifications: Models often involve simplifications to make them tractable. It's essential to understand what assumptions are made in this model and whether they adequately represent the real-world financial system. Causality: The abstract mentions a correlation between securities market liberalization and a drop in relative liquidity. However, establishing causality is challenging, and more rigorous analysis would be needed to determine whether one causes the other. Applicability: The model's predictions may be highly context-dependent, and the abstract doesn't discuss the extent to which the model's findings can be generalized to different financial systems, regions, or economic conditions.

8. <u>A Siamese network framework for bank intelligent Q&A prediction</u> Wei Wei Yingli Liang *Journal of Forecasting (2022)*	With the development of financial technologies and the explosive growth of data, the intelligent customer services have attracted a considerable attention of academic and industrial experts. The calculation of question similarity has become a key to the question and answer (Q&A) of financial intelligent customer service. In this work, we propose a Siamese network called W2V-Siamese-BiLSTM for bank Q&A prediction. Based on the Word2vec text information vectorization, the pre-training model of embedding layer is established, and two bidirectional long short-term memory (Bi-LSTM) networks are introduced to encode the upper layer input. The weights are shared in the encoding layer, and finally the Manhattan distance is used in the similarity calculation layer. The experimental results show that the supervised similarity calculation framework proposed in this work has good applicability in real financial Q&A. The accuracy of the proposed method is 81.2%.	Innovative Approach: The abstract presents an innovative approach, using a Siamese network called W2V-Siamese-BiLSTM, for question similarity calculation in financial intelligent customer service. This approach incorporates Word2vec for text vectorization and bidirectional LSTM networks, which demonstrates a technical sophistication. Shared Weights: The use of shared weights in the encoding layer of the Siamese network can contribute to efficient model training and potentially better generalization. High Accuracy: The abstract reports a high accuracy rate of 81.2% for the proposed method, which suggests that the model is effective in predicting question similarity in financial Q&A. Practical Applicability: The abstract highlights the applicability of the proposed framework in real financial Q&A scenarios, indicating that it has potential real-world utility.	Lack of Details: The abstract provides an overview of the methodology but lacks specific technical details. This makes it difficult for readers to understand the inner workings of the proposed model and its limitations. Evaluation Metrics: While the abstract mentions high accuracy, it does not discuss other important evaluation metrics commonly used in machine learning, such as precision, recall, or F1-score. A comprehensive evaluation would require a more in-depth analysis. Generalizability: The effectiveness of the proposed model may be context-specific to financial Q&A. The abstract does not discuss the generalizability of the approach to other domains or industries. Data and Dataset: The quality and size of the dataset used for training and testing the model can significantly impact its performance. The abstract does not provide information about the dataset, which is crucial for assessing the robustness of the approach. Comparison with Existing Methods: It would be valuable to compare the proposed W2V-Siamese-BiLSTM approach with existing methods or benchmarks in the field to demonstrate its superiority convincingly.

9. Predicting corporate bankruptcy: What matters? Leon Li Robert Faff *International Review of Economics and Finance (2019)*	Whether accounting: or market-based information should be employed to predict corporate default is a long-standing debate in finance research. Incorporating a regime-switching mechanism, we establish a hybrid bankruptcy prediction model with non-uniform loadings in both accounting- and market-based approaches to reexamine the issue. We find the following. Creditors should increase the loading on market-based information when large and liquid corporations are considered. Conversely, for companies with incremental information involved in accounting reporting proxied by discretionary accruals, banks could emphasize accounting ratio-based variables more than they are already emphasized. Since managerial discretion in accounting numbers could serve as a tool to bring undisclosed information about the firm to the public, the weight on accounting-based information could be increased for firms with high information asymmetry. In addition, the loading on market-based (accounting-based) information should be increased (decreased) during periods of financial crisis, defined by negative gross domestic product growth.	Revisiting a Long-standing Debate: The research addresses a longstanding debate in finance research regarding the use of accounting-based versus market-based information for predicting corporate default. By revisiting this debate, the research contributes to the ongoing discussion in the field. Regime-Switching Mechanism: The research incorporates a regime-switching mechanism, which allows for a more nuanced analysis by considering different economic conditions or periods. This adds complexity and realism to the model, potentially improving its predictive power. Hybrid Model: The research proposes a hybrid bankruptcy prediction model that combines accounting- and market-based information. Such a model has the potential to capture a broader range of factors influencing corporate default, leading to more accurate predictions. Practical Guidance: The findings provide practical guidance for creditors and banks on how to adjust their reliance on accounting- and market-based information depending on the characteristics of the corporations they are assessing and the economic environment. Consideration of Information Asymmetry: The research considers the role of information asymmetry and suggests that accounting-based information may be more valuable for firms with high information asymmetry.	Complexity: The incorporation of a regime-switching mechanism and the hybrid model approach can introduce complexity to the analysis. This complexity may make it challenging for practitioners to implement the model effectively. Data and Calibration: The accuracy of the model's predictions may be highly dependent on the quality and availability of data and the appropriate calibration of the regime-switching mechanism. Any shortcomings in data quality or calibration can affect the reliability of the findings. Interpretability: The model's results may be difficult to interpret for non-experts in finance and economics. It could benefit from providing more straightforward guidelines or recommendations for practitioners. Generalizability: The findings may be specific to the particular dataset and economic conditions considered in the research. The generalizability of the model to other contexts or time periods is a potential limitation. Assumptions: The model likely relies on certain assumptions and simplifications, as is common in quantitative finance research. These assumptions should be clearly stated, and their implications should be considered when interpreting the results.

10. ANALISIS PENGARUH RASIO KEUANGAN RBBR TERHADAP PERTUMBUHAN LABA BANK (STUDI KASUS PT. BANK CENTRAL ASIA (BCA), Tbk) Hana Tamara Putri	This research aims to analyze the effect of the NPL, LDR, ROA, ROE, NIM and CAR variables toward Earnings Growth.The Data was used in this research based on publicity annual report of PT. Bank BCA, Tbkgot from website of Bank BCAsince 2004until 2014. During research period show as data research was normally distributed. Based on multicollinierity test, heteroscedasticity test, and autocorrelation test variable digressing of classic assumption has not founded, which indicate that the available data has fulfill the condition to use multilinier regression model. Empirical evidence show as NPLandROA partiality havean influence toward earnings growth with significance value less than 0.05.ROE, LDR, NIM, and CAR variables have no influence toward earnings growth at significance level 5%. Prediction capability from these ten variables toward earnings growth is 46.1%,where the balance (53.9%) is affected to other factors which was not to be entered to this research model.Keywords : NPL, LDR, ROA, ROE, NIM, CAR, Earnings Growth	Statistical Testing: The research conducts various statistical tests, including tests for multicollinearity, heteroscedasticity, and autocorrelation, to assess the validity of the regression model and the data's suitability for analysis. This demonstrates a commitment to robust statistical analysis. Partial Effects: The research identifies which variables (NPL and ROA) have a significant influence on earnings growth, which can provide specific insights into what factors are driving or impeding earnings growth. Prediction Capability: The research quantifies the prediction capability of the model, indicating that it can explain 46.1% of the variation in earnings growth. This information helps in understanding the model's explanatory power	Limited Scope: The research focuses on a single bank (PT. Bank BCA) over a specific time period (2004-2014). While this may provide valuable insights into this particular case, the findings may not be generalizable to other banks or different time periods. Omitted Variables: The research acknowledges that 53.9% of the variation in earnings growth is affected by factors not included in the model. This highlights the presence of omitted variables, which could potentially confound the results. Causality: The research analyzes correlations between variables but does not establish causality. For example, while it finds correlations between NPL, ROA, and earnings growth, it does not explain the causal mechanisms underlying these relationships. Data Source: The data is sourced from annual reports of PT. Bank BCA, which may have limitations in terms of completeness and accuracy. Additionally, the research does not discuss any potential data issues or biases.

Review on Papers related to Prediction of Bank failures due to the Certain Bank operations

11. Bootstrap-DEA management efficiency and early prediction of bank failure: Evidence from 2008-2009 U.S. bank failures	This paper examines prediction of U.S. bank failure with a probit model that uses bias-corrected technical efficiency estimated using bootstrap data envelopment analysis as the measure of management quality. The model is tested on a sample of failed and non-failed banks during the sub-prime mortgage meltdown, 2008–2009. This measure of managerial quality allows more accurate prediction of failure than other measures. The model successfully predicts bank failure one and two years prior to failure.	This model is the first to use the bootstrap DEA efficiency score for its advantage over the Simar and Wilson (1998), DEA score. The model is tested using independent variables from one year before failure, and from two years before failure. Results for both tests are provided**.**	Failed banks have on average relatively less equity, lower management efficiency, lower earnings, and a higher proportion of nonperforming loans than non-failed banks. One year prior to failure, failed banks are less liquid than non-failed banks. However, two years before failure, failed banks are more liquid than non-failed banks. The mean of each variable except the bias corrected managerial efficiency scores
12. China's commercial bank stock price prediction using a novel K-means-LSTM hybrid approach	China's commercial Bank shares have become the backbone of the capital market. The prediction of a bank's stock price has been a hot topic in the investment field. However, the stock price is always unstable and non-linear, challenging the traditional statistical models. Inspired by this problem, a novel hybrid deep learning approach is proposed to improve prediction performance. By modifying the distance measurement algorithm into DTW, an improved K-means clustering algorithm is proposed to cluster out banks with similar price trends.	Deep learning also has unique advantages in processing unstructured data such as natural language and has a powerful ability to mine emotional information and analyze public opinion. Therefore, natural language processing technology will be used to analyze market sentiment to improve prediction accuracy. The high-frequency data will also be used to improve the accuracy of the forecast. On the other hand, the performance of long-term prediction cannot achieve an outstanding level	The statistical characteristics of BCM, CEB, BOC, and CNCB, such as the difference in mean and standard deviation, indicate different volatility. The significant difference of the ADF test indicates that their stability is also different.

13. Computational fluid dynamics predictions of critical hydrodynamics thresholds in the erodibility of inland waterway bank by ship-induced waves	A three-dimensional numerical simulation method based on the resolution of Navier–Stokes equations is used to determine the causal relationship between the ship-inducedwaves and the erosion phenomenon of the inlandwaterway banks, and alsoto provide the critical conditions for waterway bank protection. The numerical results show that by approaching the peak in the shear stress on the waterway bank.	The shipinduced waves characteristics that control sediment rates when the shoreline is attacked remain unknown or incomplete. Hence this study is based on the CFD formulation which make it possible to highlight the relationship that exists between the ship velocity, ship draft, the geometric characteristics both of the ship and waterway.	In order to better examine the erosion processes of the banks, a coupled morphodynamical model can be added to numerically restore the sediment erosion, discharge and deposition behaviors and their relative relationships under the action of shipinduced waves
14. Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach	We develop early warning models for financial crisis prediction applying machine learning techniques on macrofinancial data for 17 countries over 1870–2016. Most nonlinear machine learning models outperform logistic regression in out-of-sample predictions and forecasting. We identify economic drivers of our machine learning models by applying a novel framework based on Shapley values, uncovering nonlinear relationships between the predictors and crisis risk	We find that it is truly the global slope which matters most rather than the slope in, for example, a dominant country in the global financial system. However, we also find a strong interplay between the global slope and recessions, suggesting that its power in predicting financial crises partially stems from its role as a good leading indicator for a global economic slowdown.	With decision trees as its base model, was shown to be useful in forecasting financial crises in 100 advanced and emerging economies between 1970 and 2017 (Casabianca et al., 2019). Tölö (2019) shows that recurrent neural networks yield better early warning models than both ordinary neural networks and logistic regression in the Macrohistory database.
15. EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks	This article provides evidence that machine learning methods are suitable for reliably predicting the failure risk of European Union-27 banks from the experiences of the past decade. It demonstrates that earnings, capital adequacy, and management capability are the strongest predictors of bank failure. Critical and relevant field research is presented in the context of economic uncertainties arising from the COVID-19 pandemic.	The resulting financial ratio structures provided multi sided relationship systems with a means to explain and estimate bank failure in different forms, which presents a considerable advantage over conventional approaches.	It is somewhat uncontestable that programming 100 rulesets is a more complicated task than calculating PF values from just one logit function. However, since the C5.0 model provides classification accuracy superior to that of logit, the decision of which model to select based on their respective advantages and disadvantages is up to the end user.

16. Genomic prediction using composite training sets is an effective method for exploiting germplasm conserved in rice gene banks	Germplasm conserved in gene banks is underutilized, owing mainly to the cost of characterization. Genomic prediction can be applied to predict the genetic merit of germplasm. Large-scale resequencing projects in rice have generated high quality genome-wide variation information for many diverse accessions, making it possible to investigate the potential of genomic prediction in rice germplasm management and exploitation	Because germplasm preserved in gene banks is highly diverse, compiling a unique training set for each specific subspecies or subpopulation is impractically expensive. A composite training set consisting of accessions from each subspecies or subpopulation offers potential for comprehensive prediction, as it is related to each subspecies or subpopulation.	This reduces the benefits from assuming different marker effect distributions by the Bayesian approaches, which is more appropriate for more homogeneous populations, or even makes the assumption a disadvantage
17. Ocean quahog (Arctica islandica) growth rate analyses of four populations from the MidAtlantic Bight and Georges Bank	Growth rates from 1480 Arctica islandica from New Jersey, collected in 2019 from north and south of the Hudson Canyon, were analyzed and compared to animals obtained from Long Island and Georges Bank. New Jersey represents the southern portion of the A. islandica stock in the Mid-Atlantic Bight, and animals here may experience warmer temperatures compared to their northern counterparts	Three growth models, von Bertalanffy, Tanaka, and modified Tanaka were examined for goodness of fit to growth data. The von Bertalanffy, commonly used in fisheries management, had the worst fit for all populations, males and females, and at all 20-year cohort groups, and should not be used in the management of this species	The von Bertalanffy model provided the worst fit for both populations, overestimating at young ages and underestimating at old ages. This outcome was first considered by Pace et al. (2017a) and subsequently confirmed by Hemeon et al., 2023 for the Georges Bank and Long Island sites. Poor fit can lead to a substantial underestimation of length
18. Predicting the revocation of a bank license using machine learning algorithms	This article presents the results of applying various machine learning methods to predict the revocation of credit organizations' licenses in Russia. The goal of the research is to predict whether the bank's license will be revoked soon. The feature space was analyzed, and additional features were calculated. Different basic classification algorithms, such as logistic regression, support vector machines classifier, decision tree, and bagging ensemble algorithm, were tested to solve the problem	An enhanced bagging-based algorithm with weighted voting was developed to improve the classification quality. The results of this research can be used both by credit organizations themselves to monitor business conditions and assess risks	A software tool in Python that allows solving problems of timely prediction of license revocation based on the developed algorithm was developed.

19. The impact of governance structure on bank performance: A cross-country panel analysis using statistical learning algorithms	This paper investigates the relationship between governance structure and bank performance in normal and crisis times. Using statistical learning algorithms on R, we regressed the profitability indicators as dependent variables on board structure, bank-specific characteristics, and macroeconomic variables to examine the impact on profitability for 76 banks over 17 countries for the period from 2007 to 2019, including normal periods and the subprime crisis time.	Macroeconomic variables to examine the impact on profitability for 76 banks over 17 countries for the period from 2007 to 2019, including normal periods and the subprime crisis time	The findings show that the duality between CEO and chair has a negative and significant impact on bank performance during the full study period. BoD size and gender diversity within the BoD have no influence on bank performance but in crisis time, a woman CEO has a positive and significant impact on bank's performance.
20. What drives the growth of shadow banks? Evidence from emerging markets	The present study analyses the factors affecting the growth of Non-Banking Financial Institutions (NBFI) and finance companies in 11 emerging market economies (EMEs) that the FSB monitors. Using data for the period 2002 to 2019 and employing the panel corrected standard errors (PCSE) method, the results indicate that the growth of banks, search for yield, demand from institutional investors, and bank regulations are the key factors affecting the growth of NBFI.	The rise of big data and fintech provide a competitive advantage to shadow banks in assessing the credit worthiness of their borrowers and offering innovative products	The models were tested for cross-section dependence and were found to be cross-sectionally dependent in most cases. The results of the above tests are presented in the appendices. Thus, panel corrected standard error (PCSE) estimation was considered appropriate to estimate the models, and the results…
21. Textual analysis of Bank of England growth forecast. Herman O. Stekler	The Bank of England publishes a quarterly Inflation Report that provides numerical forecasts and a text discussion of its assessment of the UK economy. Previous research has evaluated the quantitative forecasts that are included in these reports, but we focus on the qualitative discussion of output growth, by using an in-sample textual analysis procedure to convert these qualitative assessments into a score for each report over the period 2005-2014.	This paper compare the scores both to real-time output growth data and to the corresponding quantitative projections published by the bank. We find that overall developments in the UK economy were represented accurately in the text of the Inflation Report.	Previous research has evaluated the quantitative forecasts that are included in these reports, but we focus on the qualitative discussion of output growth.

22. Bank competition, financial development and macroeconomic stability: Empirical evidence from emerging economies Habib hussain khan	In this study, we investigate the potential contribution of bank competition to macroeconomic stability, and the interactive role of financial development. Applying a two-step dynamic panel system (GMM) to macroeconomic data from 48 developing nations from 1999 to 2018, we find a bell-shaped relationship between bank competition and macroeconomic stability.	There is an optimal level of bank competition beyond which it may foster economic and financial instability. Moreover, financial development enhances bank competition's positive impact on macroeconomic stability.	Economic stability is represented by the volatility of actual and unexpected output growth, whereas financial stability is assessed by the aggregate Z-score and volatility of the private credit-to-gross domestic product ratio.
23. The internationalization of domestic banks and the credit channel of monetary policy. Daniel Osario	Banks with a large international presence tend to tolerate more their credit risk exposition relative to domestic banks. Moreover, international banks tend to rely more on foreign funding when policy rates change, allowing them to insulate better the monetary policy changes from their credit supply than domestic banks.	We also show that macroprudential FX regulation reduces banks with high FX exposition access to foreign funding, ultimately contributing to monetary policy transmission. Overall, our results suggest that the internationalization of banks lowers the potency of the bank lending channel.	Using bank-firm loan-level data, we find that loan growth and loan rates from international banks respond less to monetary policy changes than domestic banks and that internationalization partially mitigates the risk-taking channel of monetary policy.

24. Uncertainty, credit and investment: Evidence from firm-bank matched data. Hyunjoon Lim	This paper studies how high uncertainty affects corporate bank loans, addressing the important issue of identification. In times of high uncertainty, firms reduce their credit demand due to delayed investments or a deterioration in credit worthiness. Simultaneously, banks are more exposed to negative shocks to their balance sheet, reducing credit supply.	Our findings suggest that larger firms may be predominantly affected by uncertainty shocks through the real option channel rather than the financial channel. In addition, our empirical findings on firm investments align with those on bank loans.	Simultaneously, banks are more exposed to negative shocks to their balance sheet, reducing credit supply. To isolate the uncertainty effect from the credit supply effect, we employ matched bank-firm loan data covering all loans extended by all financial intermediaries to listed firms in Korea, a bank-centered economy.

25. Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach Andreas joseph	We develop early warning models for financial crisis prediction applying machine learning techniques on macrofinancial data for 17 countries over 1870-2016. Most nonlinear machine learning models outperform logistic regression in out-of-sample predictions and forecasting.	Throughout, the most important predictors are credit growth and the slope of the yield curve, both domestically and globally. A flat or inverted yield curve is of most concern when nominal interest rates are low and credit growth is high.	We identify economic drivers of our machine learning models by applying a novel framework based on Shapley values, uncovering nonlinear relationships between the predictors and crisis risk.
26. Assessing the world Bank's growth forecast Yoichi Tsuchiya	In this study, the performance and rationality of the gross domestic product growth forecasts by the World Bank (WB) for six regional aggregates and 130 individual countries between 1999 and 2019 are assessed. A large body of literature examines macroeconomic forecasts for advanced economies by intergovernmental agencies. However, evaluations of WB forecasts for emerging and developing economies rarely exist.	The extent of improvement in forecast performance, conservativeness, optimism, and under- or over-reaction is not strongly associated with regions, exports and/or imports patterns, or income levels (with some exceptions). Uncertainty measures attached to point forecasts could provide useful information to policy makers and businesses worldwide.	Therefore, this study provides the first comprehensive investigation of the W's growth forecast only for regional aggregates not for 130 countries between 1999 and 2019.

27. Finance and productivity growth: Firm-level evidence. Oliver Levine	The effect of financing frictions on firm productivity growth is not well understood. Using a model we show that a rise in financial frictions leads to increased sensitivity of productivity growth to the use of external finance. We test this prediction using a large dataset of mostly private European firms and find strong evidence supporting the prediction.	Our findings demonstrate an important link between financial markets and the real economy, and help to explain why economic activity remains persistently depressed following financial crises.	We test this prediction using a large dataset of mostly private European firms and find strong evidence supporting the prediction.The effect of financing frictions on firm productivity growth is not well understood.

28. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. Pedro Carmona	Banks play a central role in developed economies. Consequently, systemic banking crises destabilize financial markets and hamper global economic growth. In this study, extreme gradient boosting was used to predict bank failure in the U.S. banking sector. Key variables were identified to anticipate and prevent bank defaults.	The findings indicate that lower values for retained earnings to average equity, pretax return on assets, and total risk-based capital ratio are associated with a higher risk of bank failure. In addition, an exceedingly high yield on earning assets increases the chance of bank financial distress.	The data, which spanned the period 2001 to 2015, consisted of annual series of 30 financial ratios for 156 U.S. national commercial banks. Identifying leading indicators of bank failure is vital to help regulators and bank managers act swiftly before distressed financial institutions reach the point of no return.

Review on Research Papers related to the Customer Churn using Machine Learning Solutions For Prediction.

29.Customer Churn Prediction Model using Explainable Machine learning	This paper addresses the challenge of predicting customer behaviour and relating existing customers in the rapidly growing digitilaztion era.where customers have more options to choose from subscription based products and services.It aimed to develop a unique customer churn prediction model using tree-based ML alogorithms. The cost of acquiring a new customer is 5 times higher than retaining an existing customer,making it crucial to address the customer churn products.	1. The paper addresses the important problem of customer churn prediction,which is a major threat across industries. The use of explainable machine learning models and the calculation of shapley values for feature combinations enhance model interpretability and transparency. 2. In the results,we can see that it achieved a predictive performance of 81.16% for customer cgurn uisng the develooed explainable model.	1. The paper does not provide a detailed discussion on the limitations or potential challenges of the proposed approach. 2. The paper does not mention the scalability or computational efficiency of the developed models, which could be important considerations in real-world applications.

30.Churn prediction for savings bank customers: A machine learning approach	1. This paper focuses on churn prediction for savings account customers using statistical and machine learning models, with an emphasis on improving predictive power considering the imbalance characteristics of customer churn rate in the data 2. The paper suggests that commercial banks can deploy the churn prediction model to avoid customer churn and retain funds, recommending customized campaigns to enhance customer satisfaction and experience. 3. Statistical and machine learning models were employed for churn prediction, including Logistic Regression, C 5.0, CHAID, ANN, XG-BOOST, Decision Tree, and Random Forest.	The paper utilizes model accuracy, AUC, Gini coefficient, and ROC curve for model comparison, ensuring a comprehensive evaluation of the predictive power of the models.It aslo identifies variables with high predictive power for churn prediction, such as customer vintage, age, average balance, occupation code, population type, average debit amount, and average number of transactions.	The paper does not provide a detailed discussion on the limitations or potential drawbacks of the machine learning approach used for churn prediction.The paper does not mention the specific techniques used for under-sampling to address the imbalance characteristics of customer churn rate in the data.it does not discuss the potential ethical implications or considerations related to deploying a customized campaign to avoid customer churn
31.Propension to customer churn in a financial institution: a machine learning approach	In this paper they have examined churn prediction of customers in the banking sector using a unique customer-level dataset from a large Brazilian bank.The researchers explored the main determinants predicting future client churn and compare the performance of various supervised machine learning algorithms.The random forests technique was better than any other models, when comparing to k-nearest neighbors, elastic net, logistic regression, and support vector machines, in several metrics.	In order to compare between algorithms the study runs a race between numerous supervised machine learning algorithms using the identical cross-validation and evaluation setup.In a number of metrics, the random forests method outperforms models like decision trees, k-nearest neighbors, elastic nets, logistic regression, and support vector machines.According to the model, there might be losses of up to 10% of the operating results reported by the biggest Brazilian banks in 2019, which would have a substantial impact on the economy. The findings emphasize the value of spending money on cross-selling and upselling tactics targeted at current clients, which can have advantageous side effects on client retention.	1.The paper does not provide a detailed explanation of the feature selection process or the rationale behind choosing the specific machine learning algorithms used in the analysis 2.The paper does not address the potential ethical implications of using customer data for churn prediction and the impact on customer privacy . 3.The paper does not discuss the potential limitations of using a single evaluation setup for comparing different machine learning algorithms .

32.Churn prediction for savings bank customers: A machine learning approach	this paper focuses on predicting customer churn in the banking sector using a voting approach of SVM and Random Forest The study uses data visualization to analyze previous databases of banking systems and predict customer churn .The QoS provided by the company is an important factor influencing customer churn and engagement .The SVM and Random Forest algorithms are used to find the most successful and accurate way to predict customer churning or retention 1. The paper achieved an accuracy score of 0.875 in predicting customer churn using the random forest model . 2. The random forest model achieved outstanding performance with 99% accuracy, 98.5% recall, and 98.5% F1-score on the Berka dataset, and 85% accuracy, 77.5% recall, and 77% F1-score on the Kaggle dataset .	1. The paper addresses the important issue of customer churn in the banking sector, which is a significant concern for banks in terms of retaining customers and maximizing profits. 2. The study utilizes machine learning algorithms, specifically SVM and Random Forest, to predict customer churn, which are known to be effective in handling complex datasets and providing accurate predictions.3.The paper achieves high accuracy scores in predicting customer churn, indicating the potential usefulness of the proposed approach in real-world banking scenarios. 4.The use of data visualization techniques helps in analyzing previous databases and gaining insights into customer churn patterns.	1. The paper does not provide a detailed discussion on the limitations or potential challenges faced during the implementation of the proposed approach. 2. The study relies on a specific dataset (Berka and Kaggle) for training and testing the models, which may limit the generalizability of the findings to other banking datasets. 3. The paper does not explore alternative machine learning algorithms or compare the proposed approach with other existing churn prediction methods, which could provide a more comprehensive evaluation of the model's performance.
33.Machine Learning Based Customer Churn Prediction in Banking	The paper compares the performance of different classifiers, including KNN, SVM, Decision Tree, and Random Forest, for customer churn prediction in the banking sector. Feature selection methods are used to identify the most relevant features for churn prediction. The Random Forest model, after oversampling, achieves higher accuracy compared to other models.	1. The study explores the likelihood of churn by analyzing customer behavior, which provides valuable insights for banks to take preventive measures against customer turnover. 2. The paper compares the performance of different machine learning models, including KNN, SVM, Decision Tree, and Random Forest, for customer churn prediction, providing a comprehensive evaluation of these models in the banking context.	1. The paper does not mention the size or representativeness of the dataset used for experimentation, which could affect the generalizability of the results. 2. The paper does not discuss the computational complexity or resource requirements of the proposed method, which could be important considerations for practical implementation.

34.Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network	1. This study proposes a method based on generative adversarial network (GAN) to deal with the problem of poor prediction performance of traditional classifiers for minority class in imbalanced bank customer data. 2. The method involves generating minority class samples using GAN and then training a classifier on the balanced data. 3. The method outperforms traditional data sampling methods such as SMOTE and BSSMOTE in terms of F1, Precision, and other indicators	1. This study proposes a method based on generative adversarial network (GAN) to deal with the problem of poor prediction performance of traditional classifiers for minority class in imbalanced bank customer data. 2. The method involves generating minority class samples using GAN and then training a classifier on the balanced data. Experimental results show that this method is feasible and applicable to the classification of imbalanced data in banks. 3. The method outperforms traditional data sampling methods such as SMOTE and BSSMOTE in terms of F1, Precision, and other indicators	1. The paper does not provide a detailed comparison with other existing methods or algorithms for dealing with imbalanced data in the banking sector. 2. The paper does not discuss the potential limitations or challenges of implementing the proposed method in real-world banking scenarios
35.Churn Prediction in Banking Sector	Churn prediction in the banking sector is important due to the competitive services offered by banks, which can lead to customer dissatisfaction and churn.The project aims to predict churn using previous banking system databases and data visualization.Customer churn and engagement have become major problems for banks, as customers are more attracted to the quality of service provided.The project uses a voting approach of SVM and Random Forest to find the most successful and accurate way to predict customer churn or retention.	1. The paper proposes the use of LSTM model and SMOTE technique for data preprocessing, which can improve the accuracy of churn prediction. 2. The paper provides empirical evidence that the proposed systems for churn prediction perform with an accuracy of 88%, which is better than the system without SMOTE technique .	1. The paper does not provide a detailed comparison with other existing churn prediction techniques, limiting the understanding of the novelty and effectiveness of the proposed approach. 2. The paper does not discuss the limitations or potential challenges of using LSTM model and SMOTE technique for churn prediction in the banking sector. 3. The paper does not provide insights into the interpretability of the churn prediction model and how the results can be practically applied in the banking sector.

36.Machine Learning Based Customer Churn Prediction in Banking	1. The study analyzes customer behavior to explore the likelihood of churn. The KNN, SVM, Decision Tree, and Random Forest classifiers are used in the study. Feature selection methods are employed to identify relevant features and verify system performance. 2. The results show that the Random Forest model, after oversampling, outperforms other models in terms of accuracy. 3. The use of the Random Forest model after oversampling is found to have higher precision and predictability compared to other models.	1. The paper addresses an important and relevant problem in the banking sector, which is predicting customer churn. It explores the use of machine learning techniques, such as KNN, SVM, Decision Tree, and Random Forest, to predict customer churn. 2. The study uses real data and conducts experiments on a churn modeling dataset from Kaggle, which enhances the validity and applicability of the findings.	1. The paper does not provide detailed information about the specific dataset used, such as the size, source, and characteristics of the data. 2. The paper does not discuss the limitations or potential biases of the machine learning models used, which could affect the generalizability of the findings. 3.The paper does not mention the specific evaluation metrics used to assess the performance of the machine learning models, making it difficult to compare the results with other studies.
37.CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS	1.The study uses various machine learning models, including k-means clustering, k-nearest neighbors, logistic regression, decision tree, random forest, and support vector machine, to predict customer churn. 2.The results show that the random forest model performs well with an accuracy of about 97%. After customer segmentation, all models perform well, with logistic regression having the lowest accuracy (87.27%) and random forest having the highest accuracy (97.25%).	1. The paper addresses an important problem in the banking sector, which is customer churn prediction. Customer churn can have negative impacts on a company's operations and profits, so developing accurate prediction models is crucial. 2. The results of the paper demonstrate the effectiveness of the random forest model in predicting customer churn, with a high accuracy rate of about 97%.	1. The paper does not discuss the limitations or potential biases of the chosen machine learning models. It would be beneficial to address any limitations or assumptions associated with the models used in the study. 2. The paper does not explore the interpretability of the models. Understanding the factors or features that contribute most to customer churn prediction could provide valuable insights for businesses to take proactive actions.
38.A HISTORY OF THE FUTURE OF BANKING: PREDICTIONS AND OUTCOMES	Provides empirical investigation on banking customers' perception of bank machines . Examines the market share of paper checks in the United States, Europe, and Japan, providing insights into payment preferences in different regions .	1.Highlights the orientation of payments in Europe towards smart cards and electronic payments, and in Japan towards cash 2.Discusses the predictions made about the growth of automated teller machines (ATMs) and electronic banking, providing historical context . Offers predictions about the growth of ACH volume and home banking, providing a basis for comparison with actual outcomes	The paper's findings and conclusions are not explicitly mentioned in the provided sources, making it difficult to assess the specific results of the study.

6.Methodology:

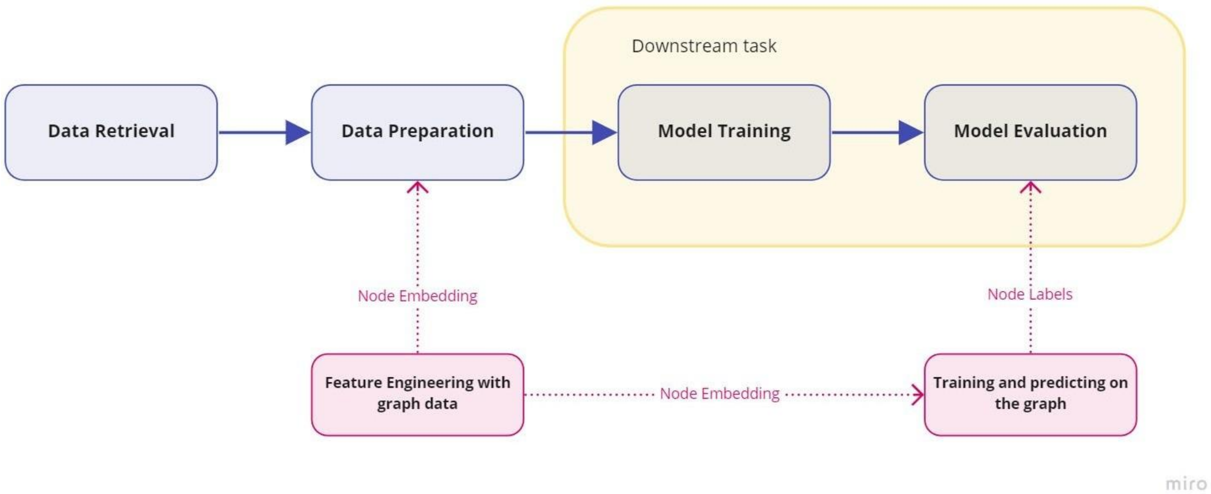
1. Data Source:
   a. By taking a glimpse at our dataset, we took a total of 10,000 and 14 columns. Three non useful variables are identified in the dataset: RowNumber, CustomerID, and Surname.
   b. Two categorical variables: Geography and Gender need to be encoded into numbers.
2. Data Processing:
   a. The data processing that need to be done include:
      i. Drop RowNumber, CustomerID, and Surname.
      ii. Encode Geography and Gender.
      iii. Log Transform Age, CreditScore, and Balance.
      iv. Scale range of Age, CreditScore, Balance, EstimatedSalary from 0 to 1.
3. Summary of the data:The summary of the statistics for the variables. When we examine the Min and Max of continuous variables, we will observe that their scales differ significantly, such as Age and EstimatedSalary. The larger scaled variables need to be scaled in the same 0 - 1 range because they will overtake the smaller scaled variables.
4. Data Analytics Models

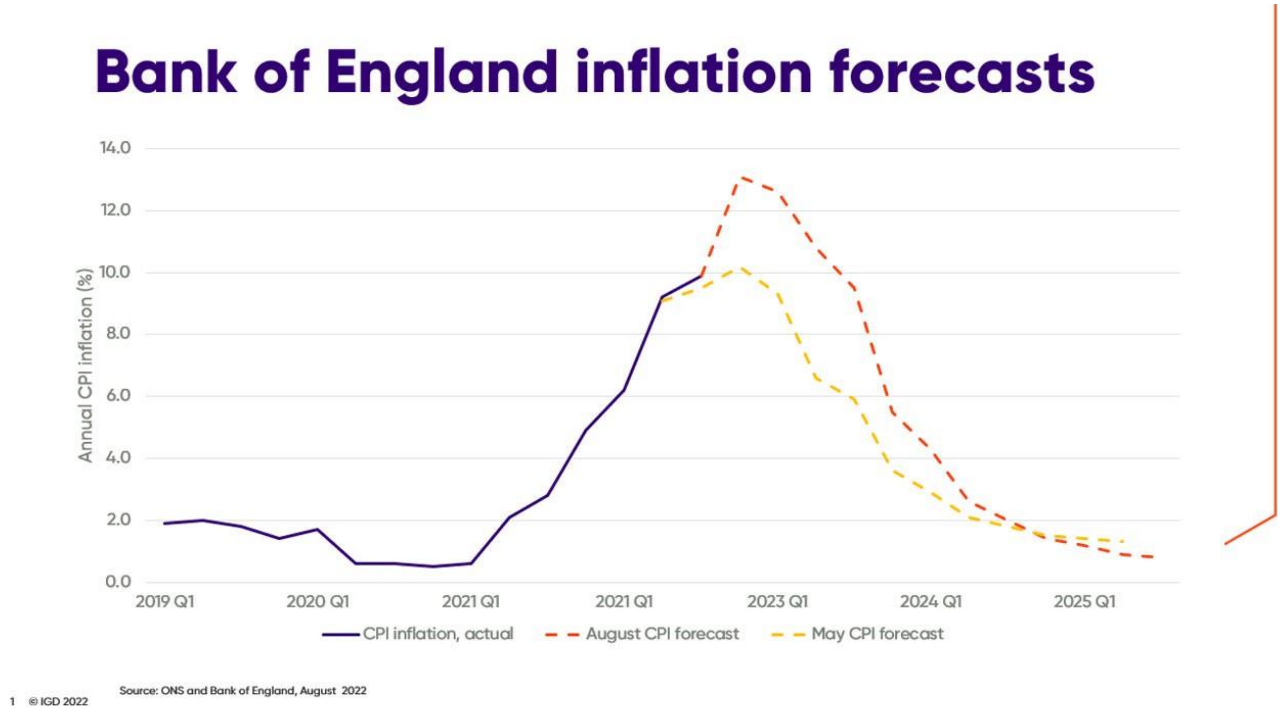
` To train a classification model, there is mainly three steps:

i. Splitting Data into Training and Testing Set
ii. Model Training/ Tuning
iii. Model Testing

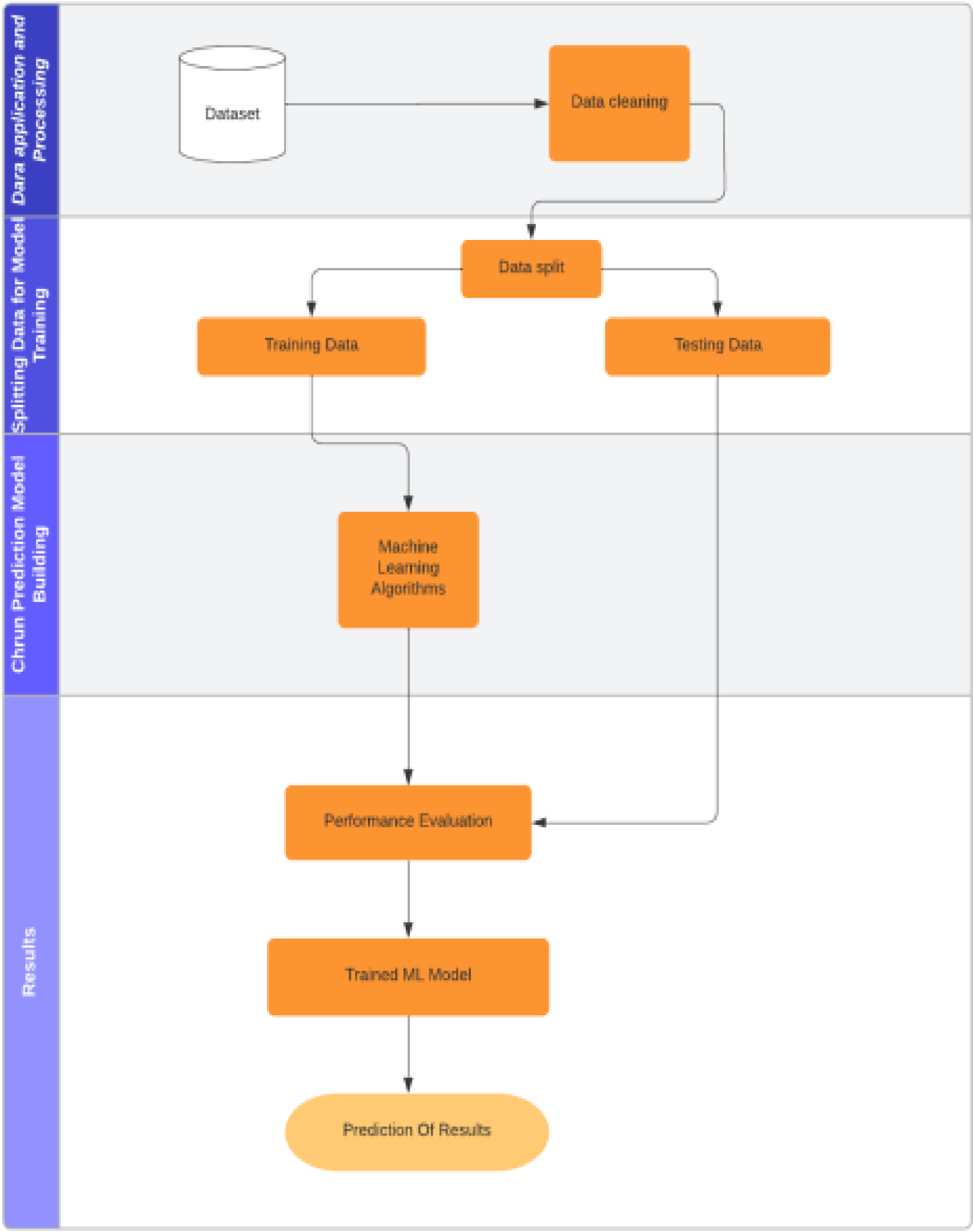
The Exited variable will be used as the target variable to predict whether a bank customer will churn or not.

Bank Growth Prediction Block Diagram and Pipeline:

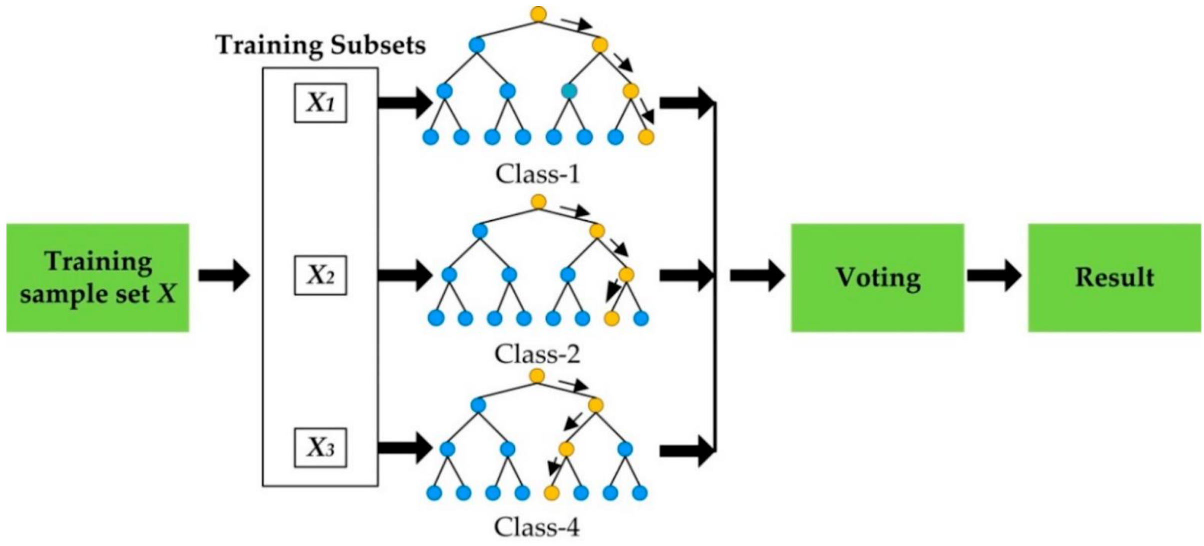




6.1. Method Approach

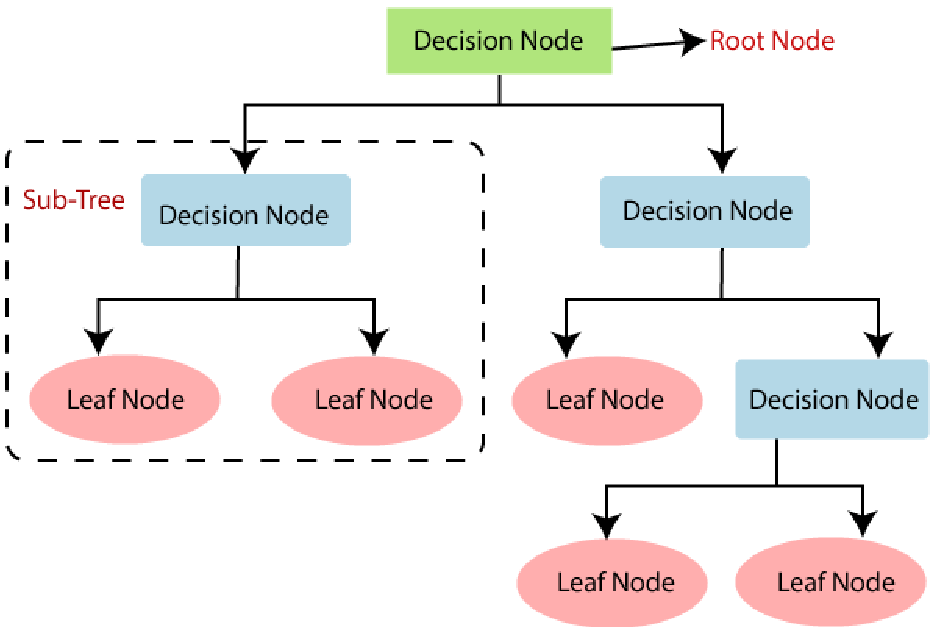

Dara application and Processing
Dataset
Data cleaning
Splitting Data for Model Training
Data split
Training Data
Testing Data
Chrun Prediction Model Building
Machine Learning Algorithms
Results
Performance Evaluation
Trained ML Model
Prediction Of Results

6.2 RANDOM FOREST CLASSIFIER



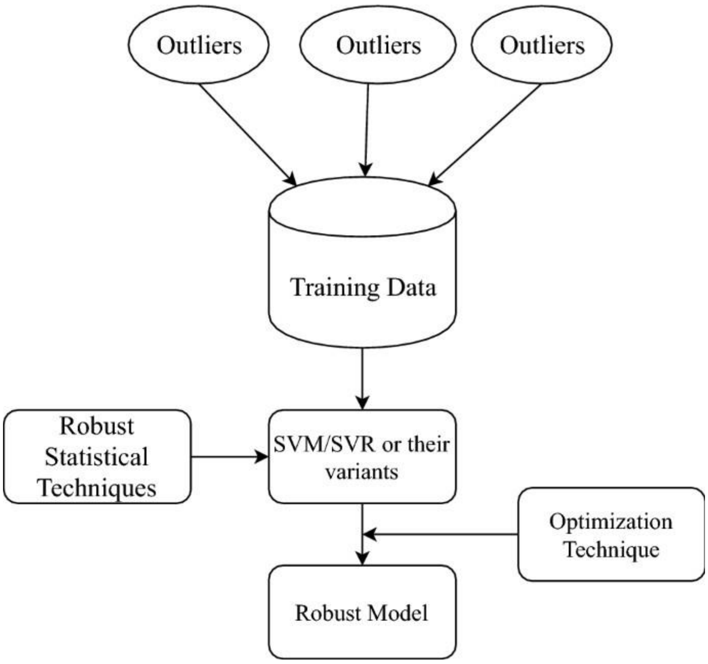
Random forest classifier is a classification technique that uses algorithms consisting of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

6.3 DECISION TREE



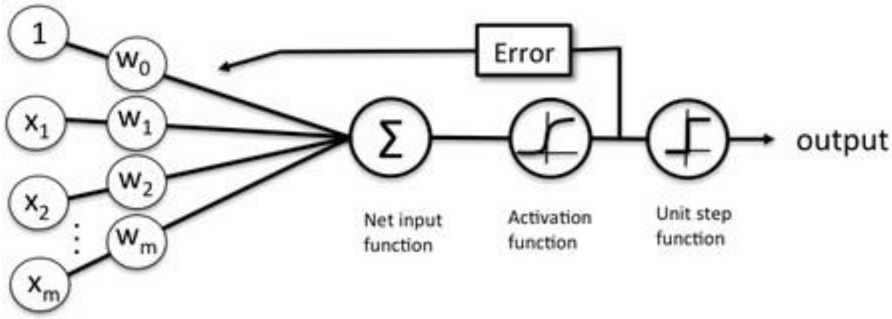
A Decision tree is a tree-like structure with attributes assigned as a node. Based on the values of the attribute's algorithm it will traverse through the tree finally ending with the leaves of the tree which contains classification output. In the Decision tree we are using the Gini index (It is a measure of the impurity of the values in the attributes and split the tree into many branches).

6.4 SVM (SUPPORT VECTOR MACHINE)



Support vector machines (SVM) are used to classify both direct and non-linear data. In short,when the algorithm receives the original training data, it uses non-linear mapping to transfigure it into an advanced dimension. In this dimension, a direct optimal hyper active airplane is sought to separate the data of any two classes. SVM can also be used for bracket and numerical validation. The simplest form of SVM is a two-class problem where the classes are linearly divisible. For a 2- D problem, a straight line can be drawn to separate the classes, in fact, multiple lines can be drawn. In the SVM algorithm, the kernels used for the classification of websites are Default, linear, RBF, and polynomial.

6.5 LOGISTIC REGRESSION

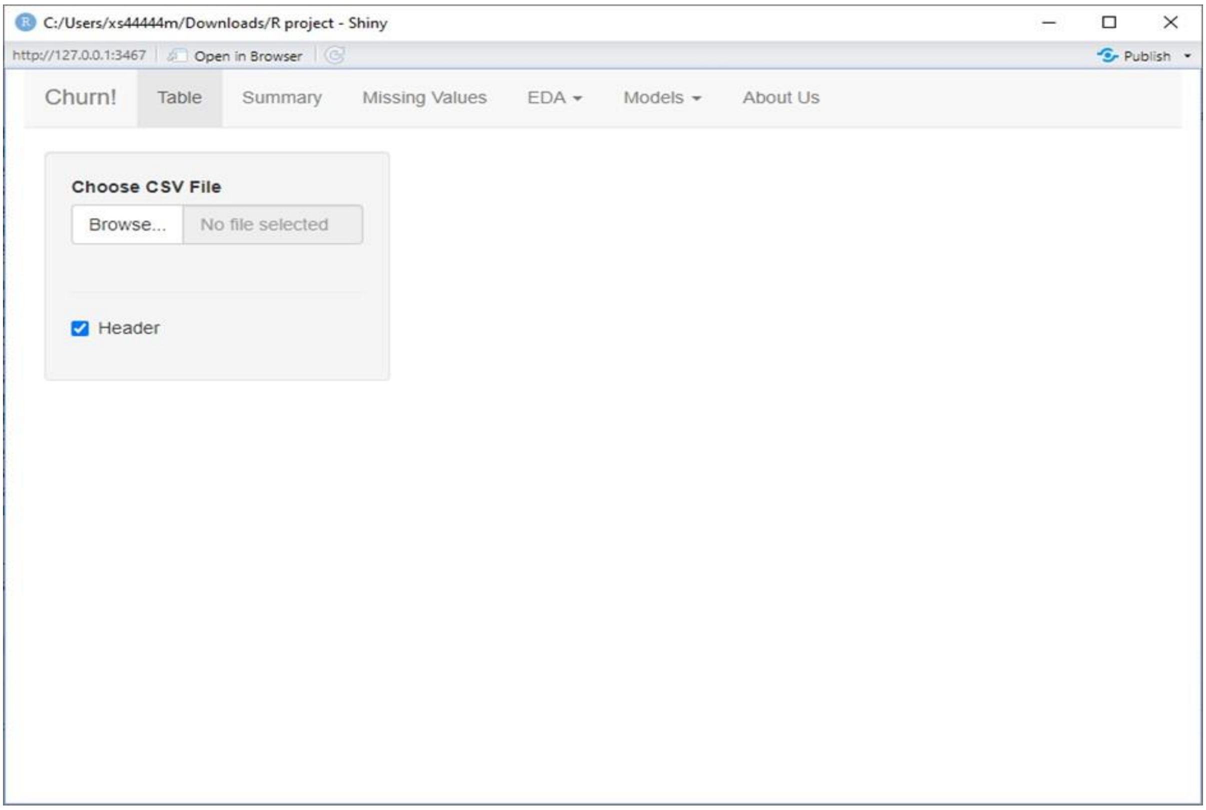


Schematic of a logistic regression classifier.

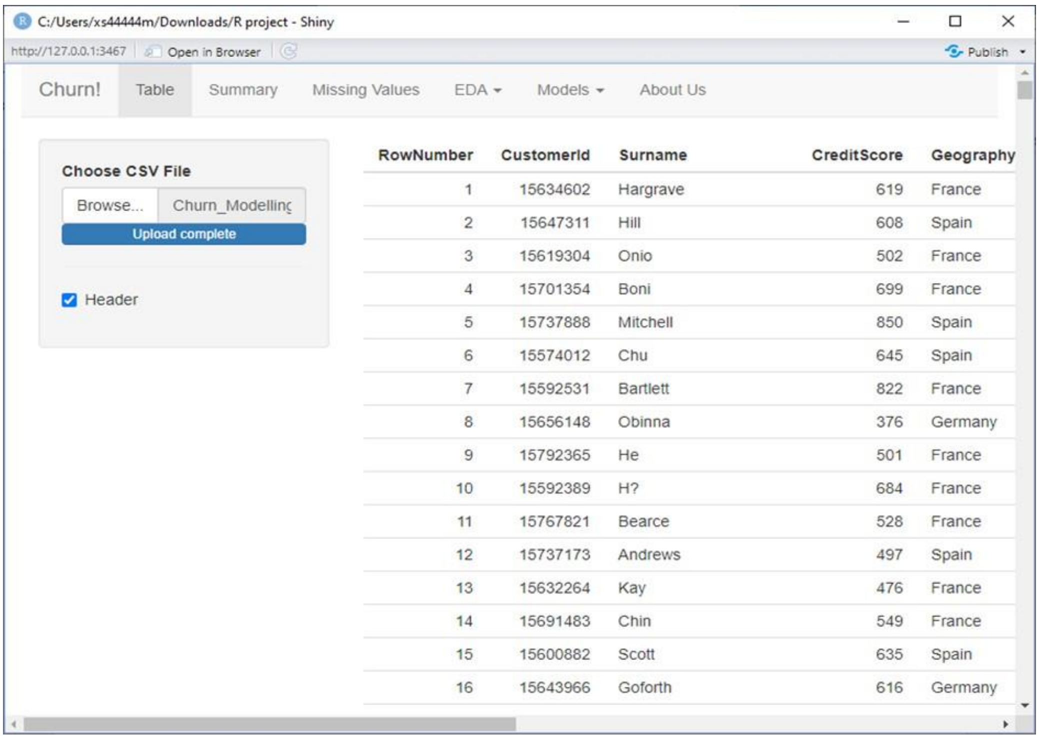
Logistic regression is a type of predictive analysis where churn prediction can be detected based on attributes. In logistic regression, the input is given as training data and test data. Based on the given input, logistic regression is calculated using a regression function called a sigmoid function,with the calculated sigmoid function, the relationship between the training data and the test data is calculated.

7.RESULT:

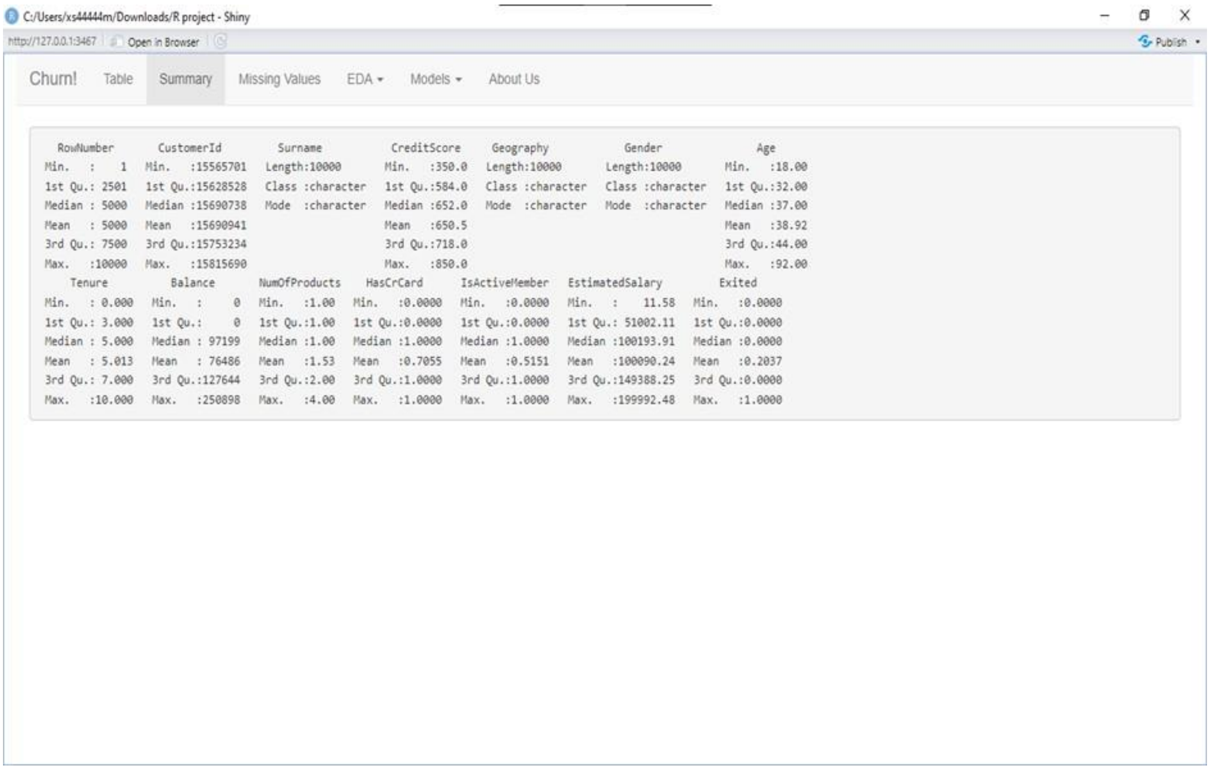
The opening webpage looks like



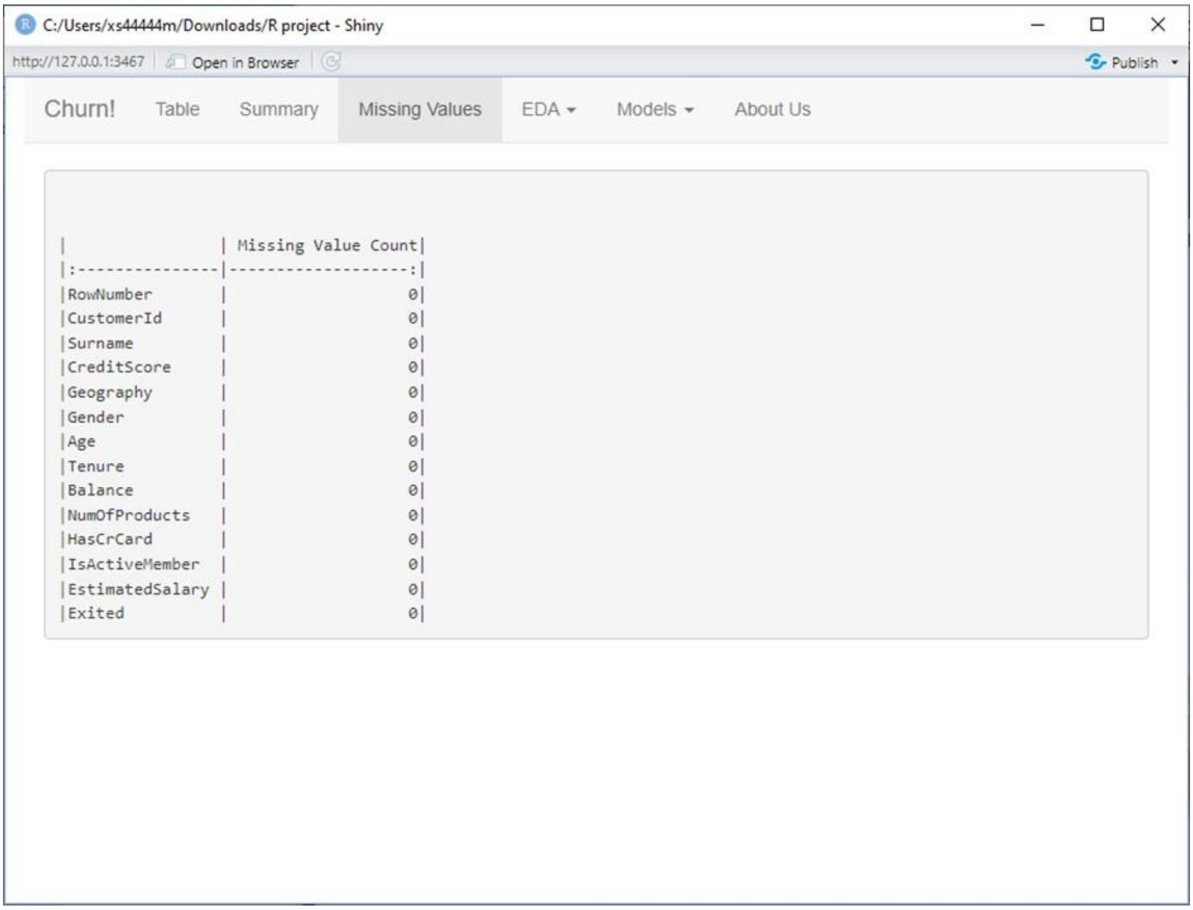
After insertion of datafile that is in csv format



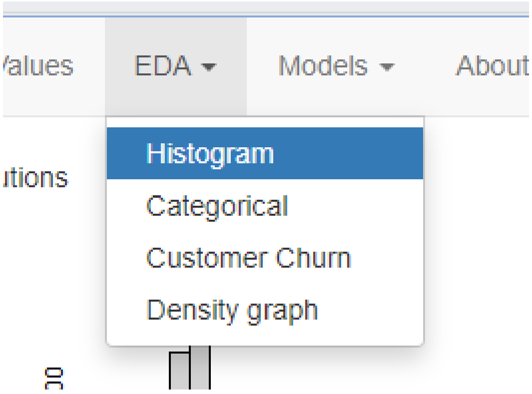
Next tab summary gives whole details of the table

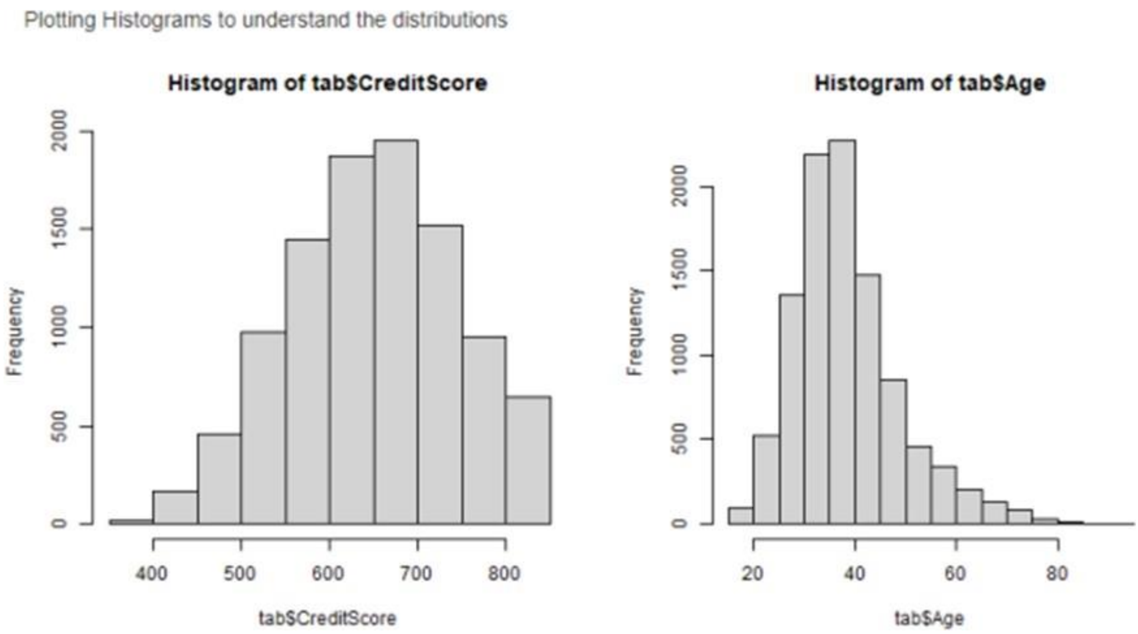


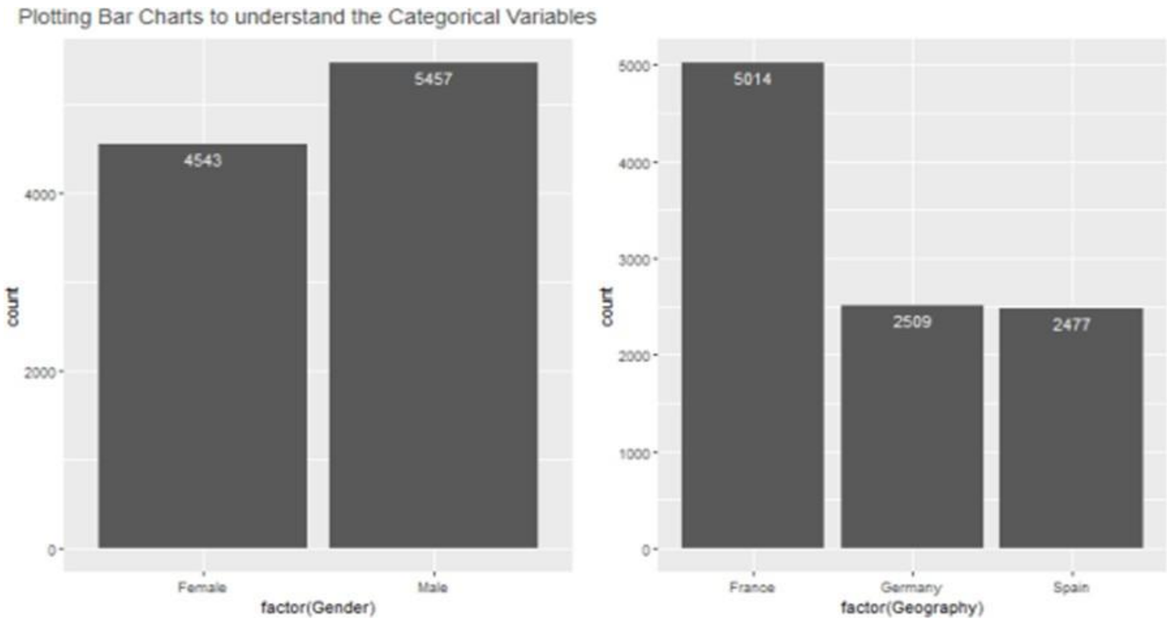
Next tab is about any missing values in the csv file

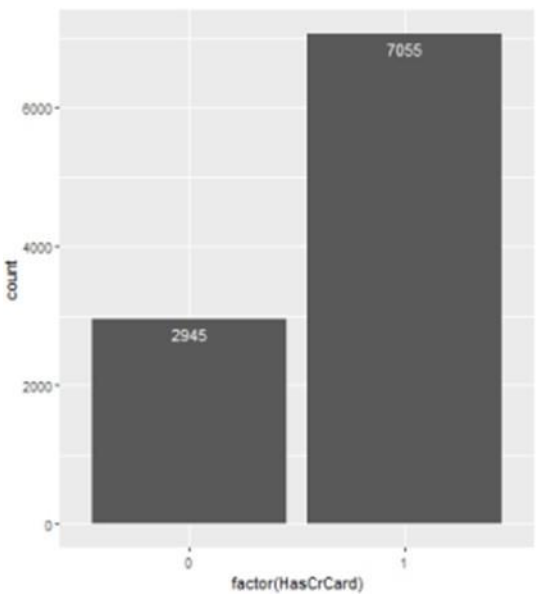


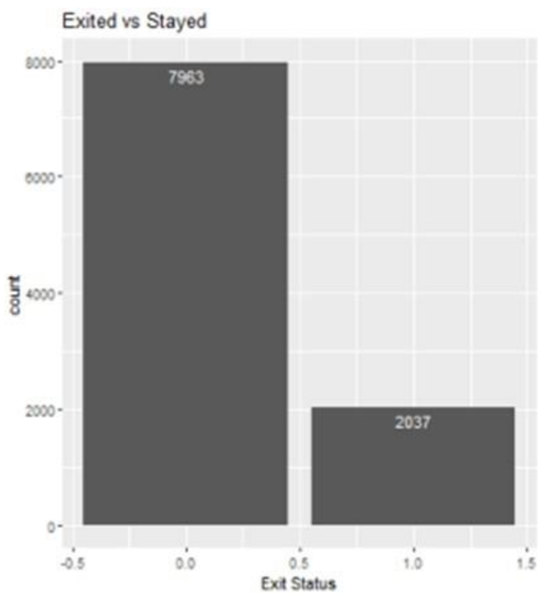
Next tab is EDA(Exploratory data analysis) where it contains different types of graphs for analysis we get options as below



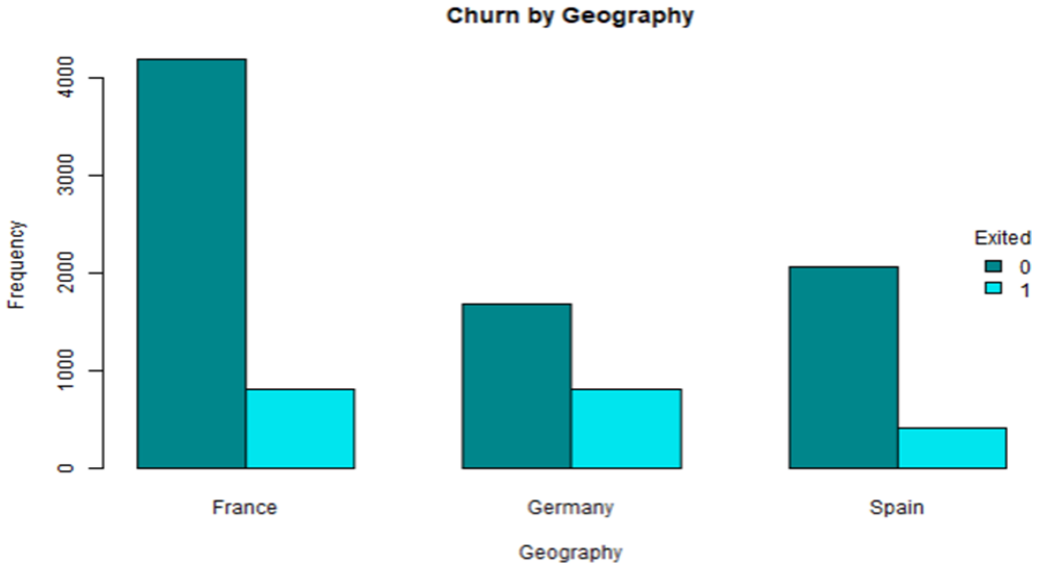


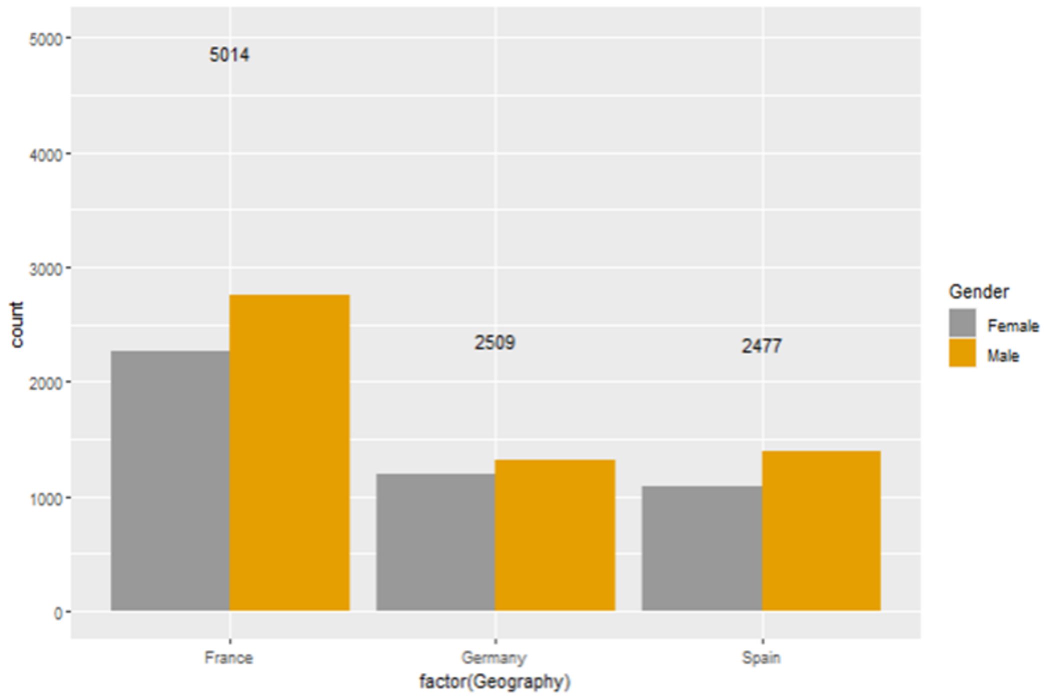

Plotting Bar Charts to understand the Categorical Variables
4543
5457
count
0
2000
4000
Female
Male
factor(Gender)
5014
2509
2477
count
0
1000
2000
3000
4000
5000
France
Germany
Spain
factor(Geography)


2945
7055
count
0
2000
4000
6000
0
1
factor(HasCrCard)


Exited vs Stayed
7963
2037
count
0
2000
4000
6000
8000
-0.5
0.0
0.5
1.0
1.5
Exit Status

Customer Churn

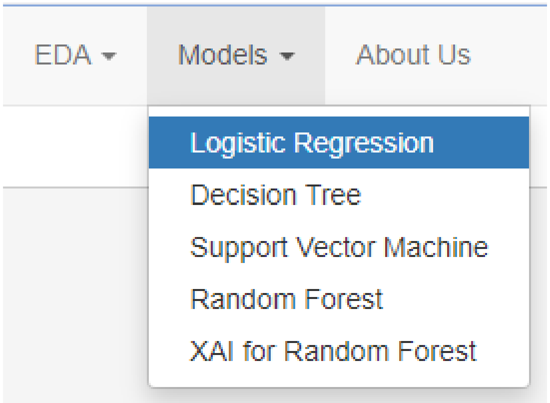


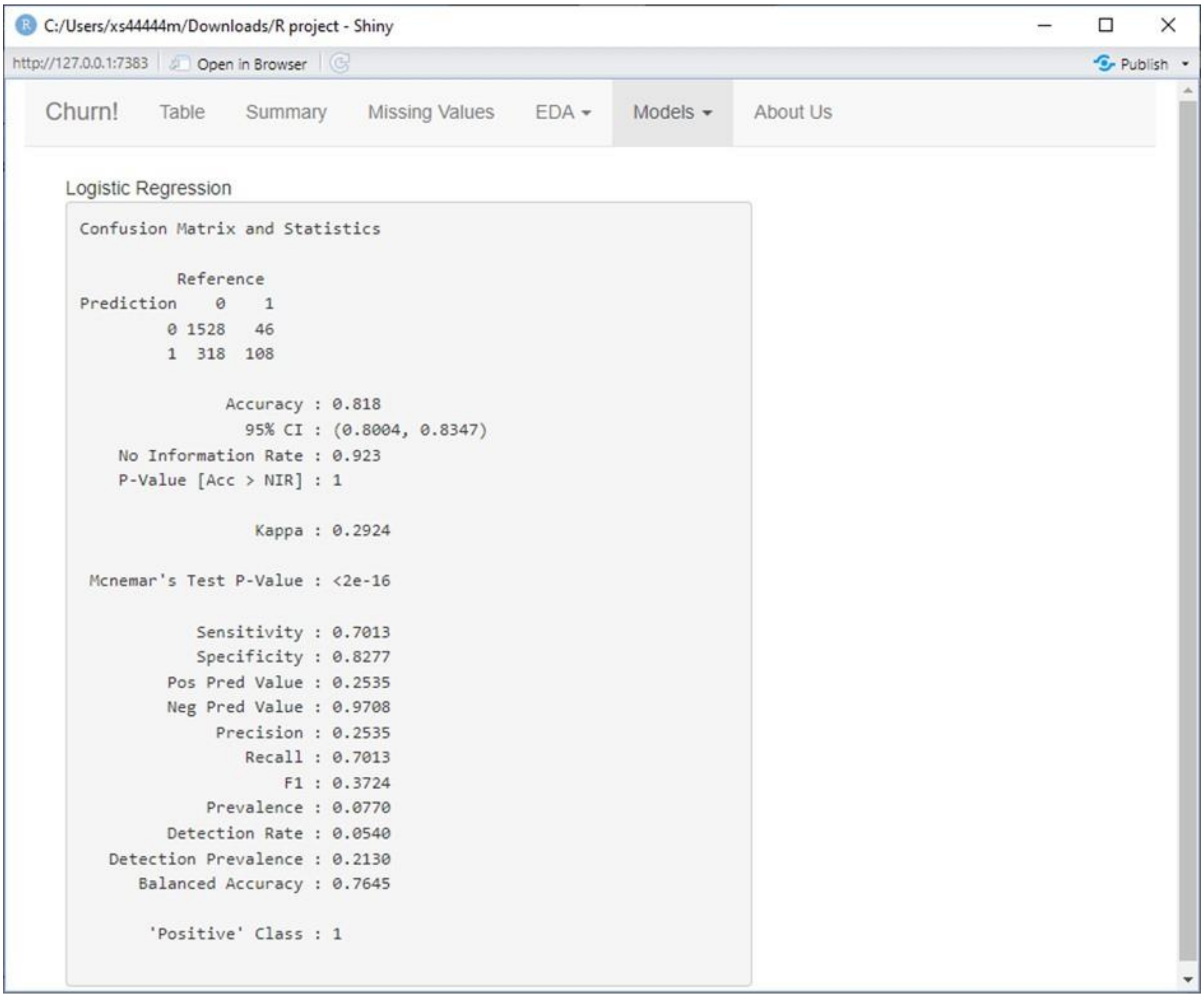


Now the Machine learning models which are done:

1.Logical Regression:

The opening of the model page look like





The logistic regression is done first, if we click on the named button, we get the page as below:

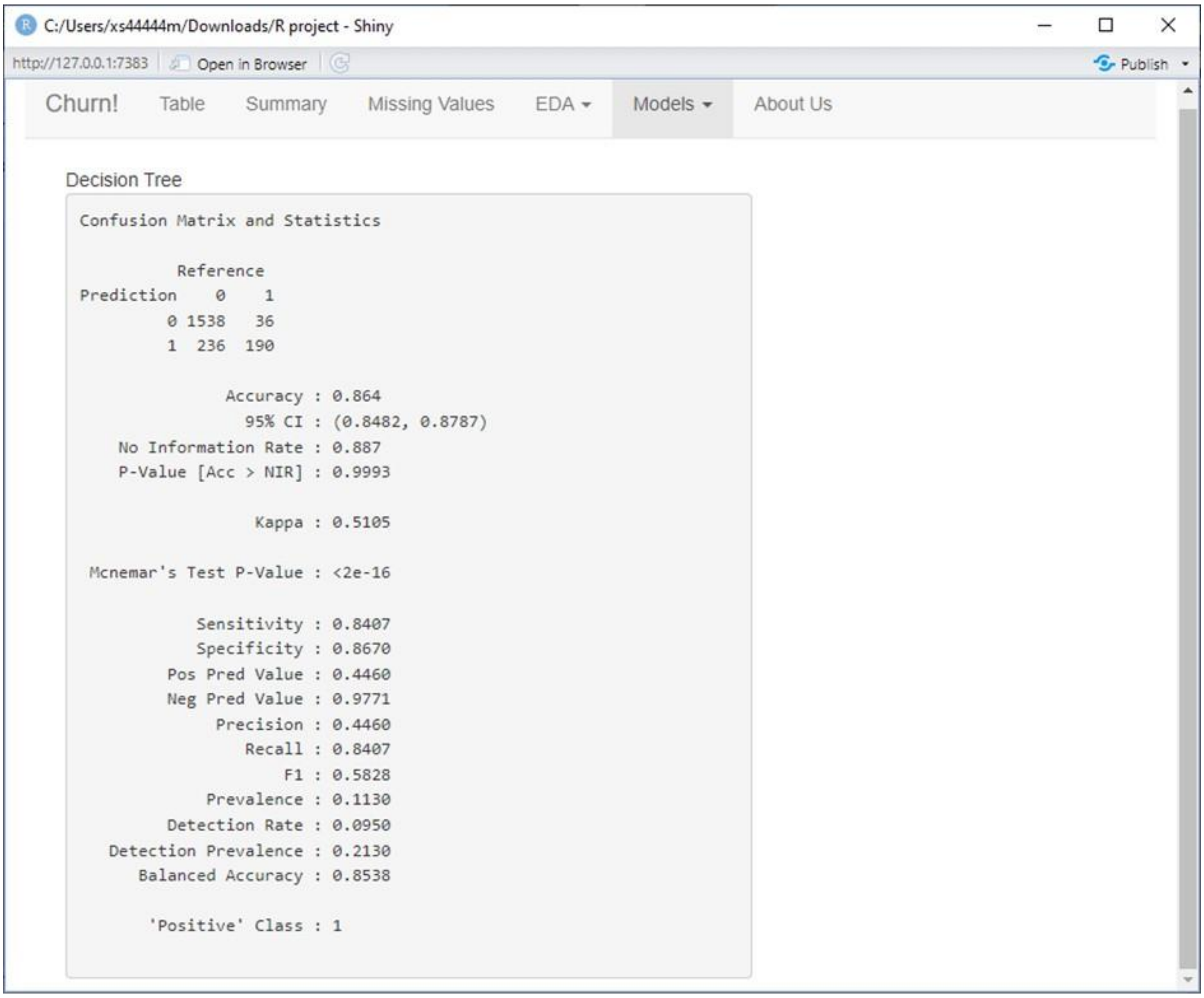
Which shows all the information of the model, and the accuracy of the model is 81.8% which is shown. The code of logical regression is:

```
output$logi_reg <- renderPrint({

        inFile <- input$file1

        if (is.null(inFile))
```

```
return(NULL)

data=read.csv(inFile$datapath, header = input$header)

data = data[, !names(data) %in% c('RowNumber', 'CustomerId', 'Surname')]

# data encoding

data$Geography = factor(data$Geography, labels=c(0, 1, 2))

data$Gender = factor(data$Gender, labels=c(0, 1))

# data transformation

data$Age = log(data$Age)

data$CreditScore = log(data$CreditScore)

data$Balance = log(data$Balance)

data[data$Balance == -Inf, 'Balance'] <- 0

# scaling

fun_scale_0to1 <- function(x) {

(x - min(x)) / (max(x) - min(x))

}

data$Age = fun_scale_0to1(data$Age)

data$CreditScore = fun_scale_0to1(data$CreditScore)

data$Balance = fun_scale_0to1(data$Balance)

data$EstimatedSalary = fun_scale_0to1(data$EstimatedSalary)

set.seed(1000)

trainIndex <- createDataPartition(data$Exited, p = 0.8, list = FALSE, times = 1)

training_data <- data[ trainIndex,]

testing_data  <- data[-trainIndex,]

LR_model = glm(Exited ~ CreditScore + Geography + Gender + Age +Tenure+ Balance +
NumOfProducts + IsActiveMember, data = training_data, family = "binomial")

pred2 <- predict(LR_model,testing_data,type="response")

cutoff_churn <- ifelse(pred2>=0.50, 1,0)

cm <- confusionMatrix(as.factor(testing_data$Exited),as.factor(cutoff_churn),mode =
"everything",positive ='1')

cm })
```

2.Decision Tree:

The page is shown as below:

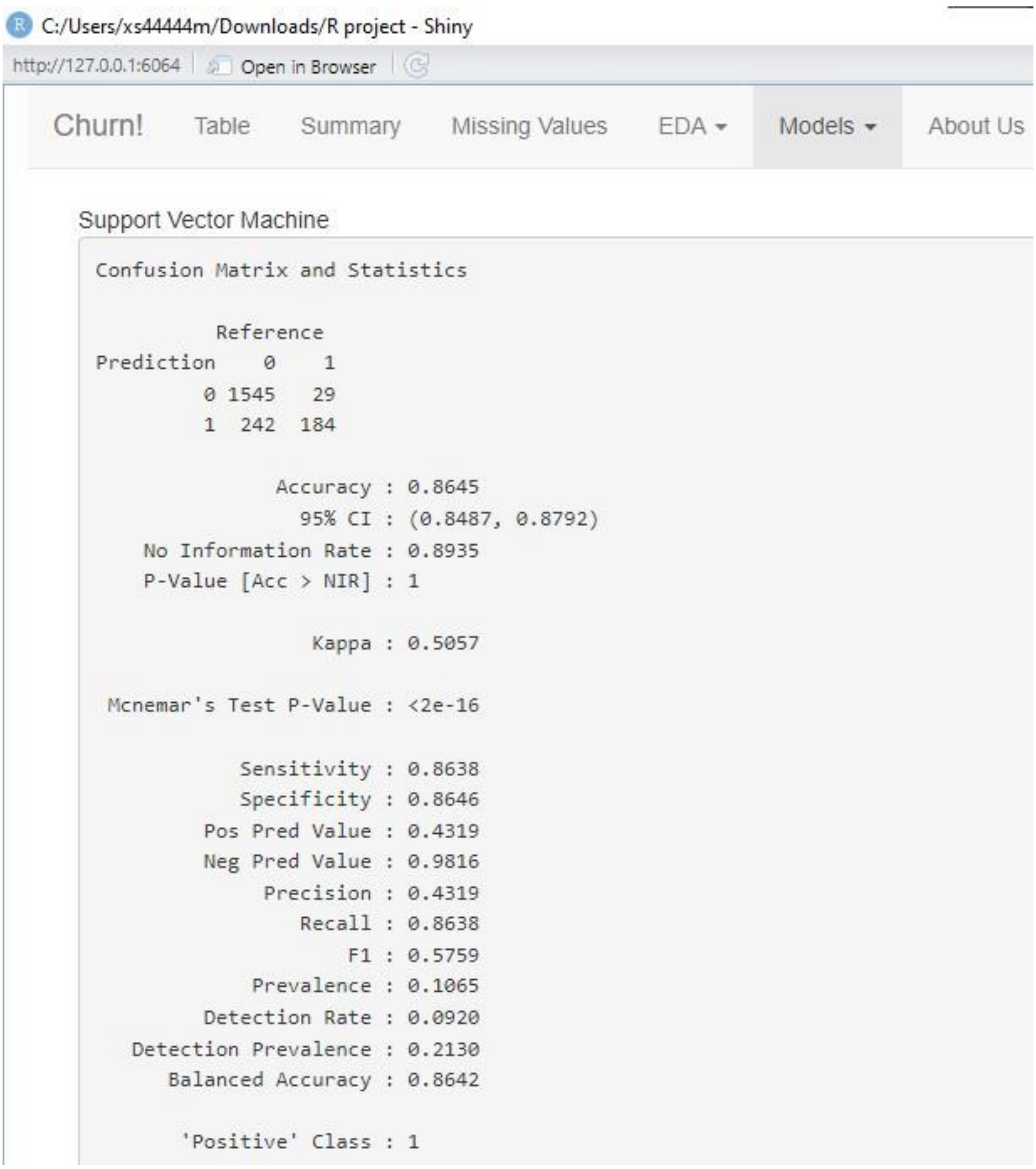


Which shows all the information of the model, and the accuracy of the model is 86.4% which is shown. The code of Decision Tree is:

```
output$dt2 <- renderPrint({

        inFile <- input$file1

        if (is.null(inFile))

        return(NULL)

        data=read.csv(inFile$datapath, header = input$header)

        data = data[, !names(data) %in% c('RowNumber', 'CustomerId', 'Surname')]

        # data encoding

        data$Geography = factor(data$Geography, labels=c(0, 1, 2))

        data$Gender = factor(data$Gender, labels=c(0, 1))
```

```
# data transformation
data$Age = log(data$Age)
data$CreditScore = log(data$CreditScore)
data$Balance = log(data$Balance)
data[data$Balance == -Inf, 'Balance'] <- 0
# scaling
fun_scale_0to1 <- function(x) {
(x - min(x)) / (max(x) - min(x))
}
data$Age = fun_scale_0to1(data$Age)
data$CreditScore = fun_scale_0to1(data$CreditScore)
data$Balance = fun_scale_0to1(data$Balance)
data$EstimatedSalary = fun_scale_0to1(data$EstimatedSalary)
set.seed(1000)
trainIndex <- createDataPartition(data$Exited, p = 0.8, list = FALSE, times = 1)
training_data <- data[ trainIndex,]
testing_data  <- data[-trainIndex,]
Dtree = rpart(Exited ~., data = training_data, method = "class")
set.seed(12345)
cv.ct <- rpart(Exited ~., data = training_data, method = "class",
        cp = 0.00001, minsplit = 5, xval = 5)
prune_dt <- prune(cv.ct,cp=cv.ct$cptable[which.min(cv.ct$cptable[,"xerror"]),"CP"])
predict_dt <- predict(prune_dt, testing_data,type="class")
cm_dt <- confusionMatrix(as.factor(testing_data$Exited),as.factor(predict_dt),mode = "everything",positive='1')
cm_dt
})
```

3.Support Vector Machine

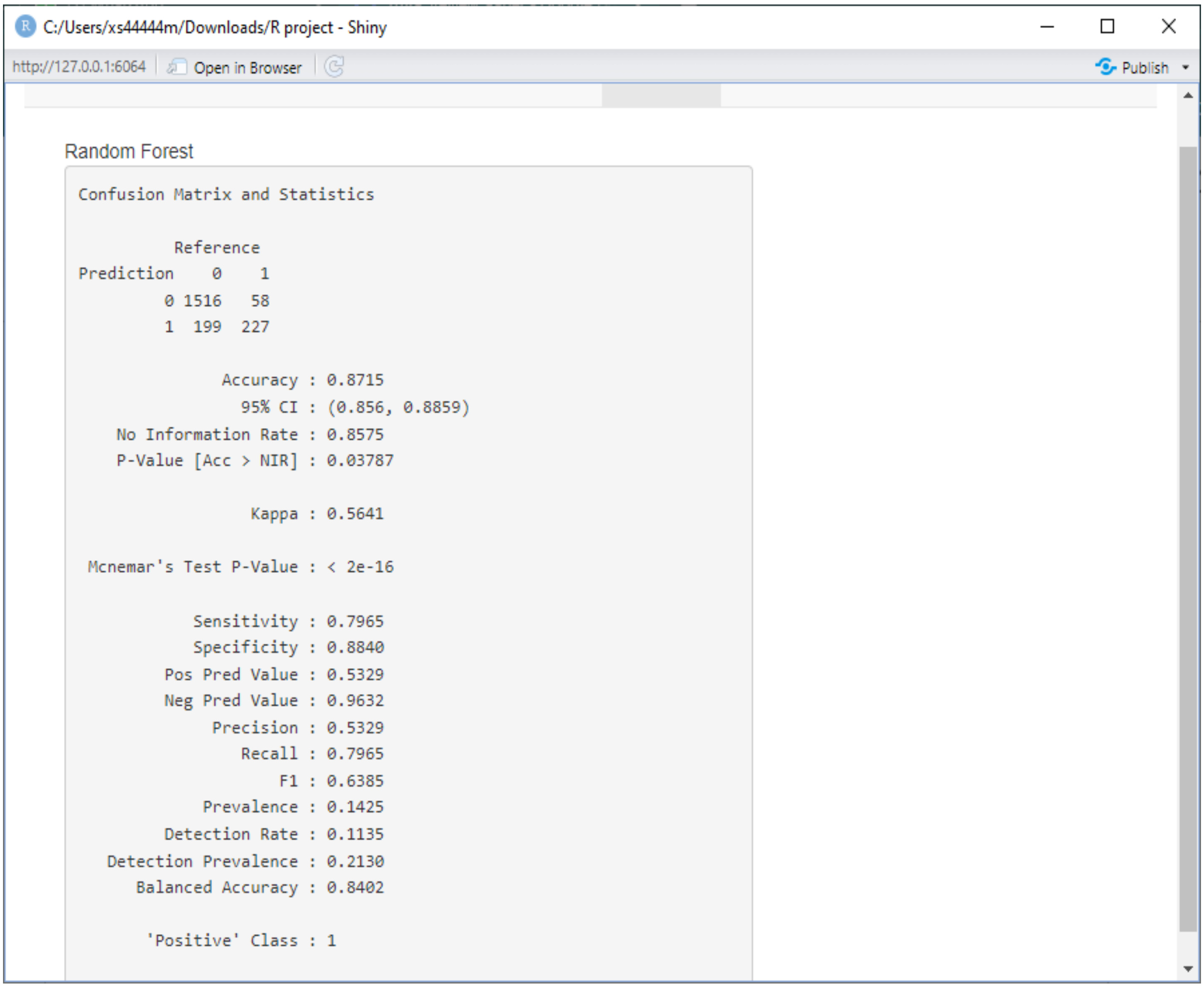


Which shows all the information of the model, and the accuracy of the model is 86.45% which is shown. The code of SVM is:

```
output$svm <- renderPrint({

  inFile <- input$file1

  if (is.null(inFile))

    return(NULL)
```

```
    data=read.csv(inFile$datapath, header = input$header)

    data = data[, !names(data) %in% c('RowNumber', 'CustomerId', 'Surname')]

    # data encoding

    data$Geography = factor(data$Geography, labels=c(0, 1, 2))

    data$Gender = factor(data$Gender, labels=c(0, 1))

    # data transformation

    data$Age = log(data$Age)

    data$CreditScore = log(data$CreditScore)

    data$Balance = log(data$Balance)

    data[data$Balance == -Inf, 'Balance'] <- 0

    # scaling

    fun_scale_0to1 <- function(x) {

      (x - min(x)) / (max(x) - min(x))

    }

    data$Age = fun_scale_0to1(data$Age)

    data$CreditScore = fun_scale_0to1(data$CreditScore)

    data$Balance = fun_scale_0to1(data$Balance)

    data$EstimatedSalary = fun_scale_0to1(data$EstimatedSalary)

    set.seed(1000)

    trainIndex <- createDataPartition(data$Exited, p = 0.8, list = FALSE, times = 1)

    training_data <- data[ trainIndex,]

    testing_data  <- data[-trainIndex,]

    learn_svm <- svm(factor(Exited)~.,data=training_data)

    predict_svm <- predict(learn_svm, testing_data,type ="response")

    cm_svm <- confusionMatrix(as.factor(testing_data$Exited),as.factor(predict_svm),mode =
"everything",positive='1')

    cm_svm

  })
```

4.Random Forest:

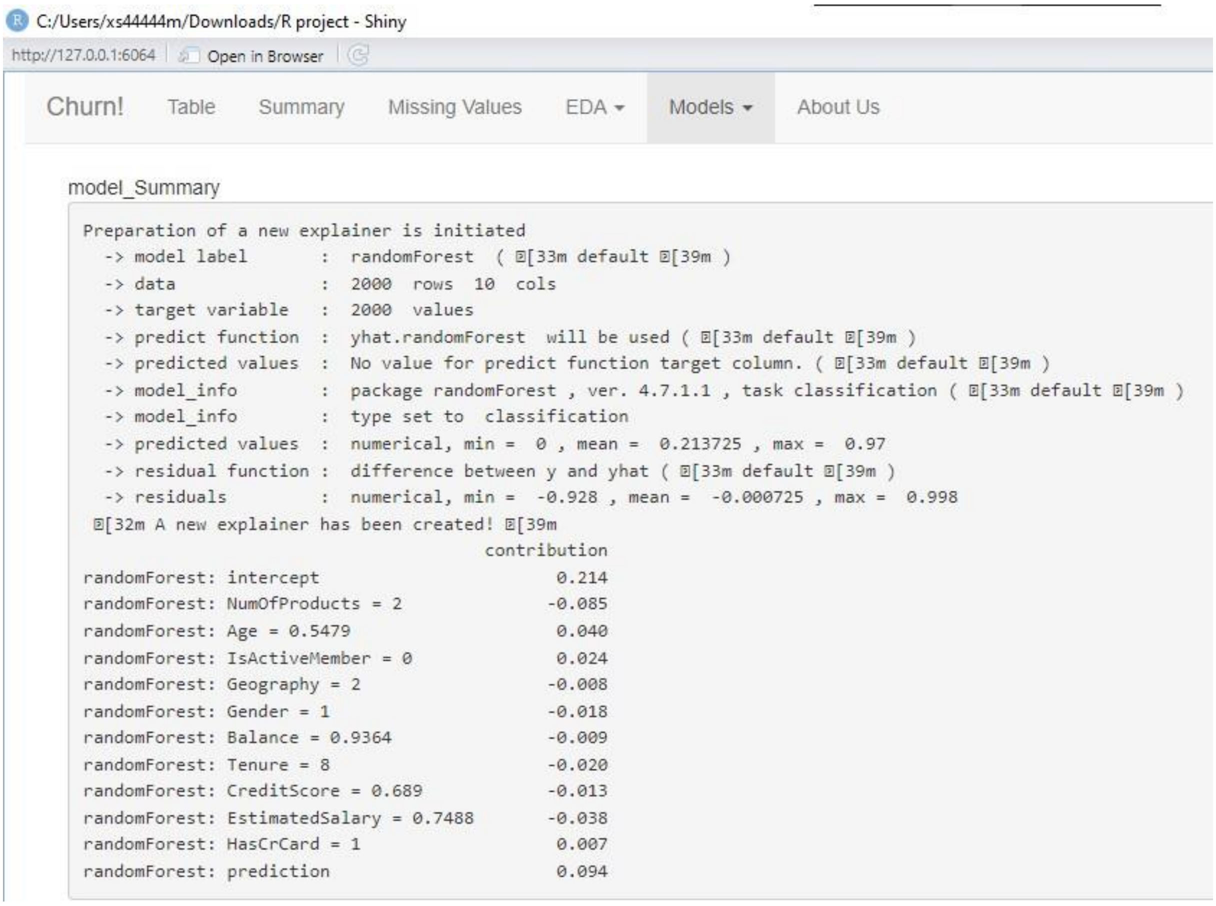


Which shows all the information of the model, and the accuracy of the model is 87.15% which is shown. The code of Random Forest is:

```
output$ran_for <- renderPrint({

   inFile <- input$file1

   if (is.null(inFile))

     return(NULL)

   data=read.csv(inFile$datapath, header = input$header)

   data = data[, !names(data) %in% c('RowNumber', 'CustomerId', 'Surname')]

   # data encoding

   data$Geography = factor(data$Geography, labels=c(0, 1, 2))

   data$Gender = factor(data$Gender, labels=c(0, 1))

   # data transformation
```

```
data$Age = log(data$Age)

data$CreditScore = log(data$CreditScore)

data$Balance = log(data$Balance)

data[data$Balance == -Inf, 'Balance'] <- 0

# scaling

fun_scale_0to1 <- function(x) {

  (x - min(x)) / (max(x) - min(x))

}

data$Age = fun_scale_0to1(data$Age)

data$CreditScore = fun_scale_0to1(data$CreditScore)

data$Balance = fun_scale_0to1(data$Balance)

data$EstimatedSalary = fun_scale_0to1(data$EstimatedSalary)

set.seed(1000)

trainIndex <- createDataPartition(data$Exited, p = 0.8, list = FALSE, times = 1)

training_data <- data[ trainIndex,]

testing_data <- data[-trainIndex,]

answer <- testing_data$Exited

rf <- randomForest(factor(Exited)~., data = training_data)

pred.rf <- predict(rf, testing_data,type ="response")

cm_rf <- confusionMatrix(as.factor(testing_data$Exited),as.factor(pred.rf),mode =
"everything",positive='1')

cm_rf

})
```

5.XAI for Random Forest:



This part explains the performance of the Random forest which shows more accuracy.The code is

```
output$ran_for_explainable <- renderPrint({

  inFile <- input$file1

  if (is.null(inFile))

    return(NULL)

  data = read.csv(inFile$datapath, header = input$header)

  data = data[, !names(data) %in% c('RowNumber', 'CustomerId', 'Surname')]

  # data encoding

  data$Geography = factor(data$Geography, labels = c(0, 1, 2))

  data$Gender = factor(data$Gender, labels = c(0, 1))

  # data transformation

  data$Age = log(data$Age)

  data$CreditScore = log(data$CreditScore)
```

```
data$Balance = log(data$Balance)

data[data$Balance == -Inf, 'Balance'] <- 0

# scaling

fun_scale_0to1 <- function(x) {

  (x - min(x)) / (max(x) - min(x))

}

data$Age = fun_scale_0to1(data$Age)

data$CreditScore = fun_scale_0to1(data$CreditScore)

data$Balance = fun_scale_0to1(data$Balance)

data$EstimatedSalary = fun_scale_0to1(data$EstimatedSalary)

set.seed(1000)

trainIndex <- createDataPartition(data$Exited, p = 0.8, list = FALSE, times = 1)

training_data <- data[trainIndex, ]

testing_data <- data[-trainIndex, ]

answer <- testing_data$Exited

rf <- randomForest(factor(Exited) ~ ., data = training_data)

pred.rf <- predict(rf, testing_data, type = "response")

cm_rf <- confusionMatrix(as.factor(testing_data$Exited), as.factor(pred.rf), mode = "everything",
positive = '1')

# Create an explainer for the Random Forest model

explainer <- DALEX::explain(model = rf, data = testing_data[, -ncol(testing_data)], y = answer, type
= "classification")

# Generate feature importance

feature_importance <- model_parts(explainer)

# Generate individual predictions explanations (e.g., SHAP values) for a specific observation

new_observation <- testing_data[1, -ncol(testing_data)] # Change this to the observation you want
to explain

predictions_explanations <- predict_parts(explainer, new_observation = new_observation)

# Create a model-level summary
```

```
   model_summary <- model_profile(explainer)

   model_summary

   predictions_explanations

 })
```

8.Discussion and conclusion:

The review paper's subject matter is of utmost importance in the financial industry, indicating a pressing concern for banks and financial institutions all over the world. The focus of this review paper is to examine the intricacies of this field, highlighting its relevance and the methodologies used to address this pressing issue.

The phenomenon of banking customer churn, or the discontinuation of customers' relationships with financial institutions, has significant financial implications. It is necessary for banks to take proactive measures to prevent the loss of valuable clients, revenue, and reputation due to high churn rates.

The paper emphasizes the crucial significance of predictive modeling, emphasizing the use of supervised learning classification algorithms to identify and comprehend the reasons for churn. The paper demonstrates how data-driven approaches can predict the likelihood of churn and provide actionable insights by analyzing a dataset of 10,000 bank records.

In addition, the discussion focuses on data preprocessing steps, including feature engineering and scaling, which are crucial for ensuring that machine learning models perform optimally. The use of algorithms like Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression demonstrates the variety of approaches employed in the field.

The importance of performance metrics, such as precision, recall, and overall accuracy, is highlighted in this review paper when evaluating the effectiveness of churn prediction models. The use of these metrics enables financial institutions to assess the reliability and efficacy of their strategies for keeping customers.

In conclusion, this review paper serves as a comprehensive resource for professionals in the banking industry, researchers, and data scientists seeking to understand and address the challenge of banking customer churn.It underscores the critical importance of data-driven predictive modeling and provides valuable insights into the methodologies and techniques employed to mitigate customer churn, ultimately contributing to improved customer retention and financial stability of banks.

References:

[1] Zhiyong Li, Chen Feng, Ying Tang. “Bank efficiency and failure prediction: a nonparametric and dynamic model based on data envelopment analysis”.

[2] Hicham Sadok, Fadi Sakka, Mohammed El Hadi El Maknouzi. “Artificial intelligence and bank credit analysis: A review”.

[3] ,Madher Ebrahim Hamdallah, Anan Fathi Srouji, *Journal of Governance and Regulation (2022)* *“*THE INFLUENCE OF SUSTAINABLE INNOVATION ON FINANCIAL ENTREPRENEURSHIP PERFORMANCE: GROWTH AND PREDICTION IN AN EMERGING MARKET”.

[4] Gabriel Chodorow-Reich, Gita Gopinath, Abhinav Narayanan. “Cash and the Economy: Evidence from India's Demonetization” .

[5] Raheela Firdaus, Yang Xue, Muhammad Sibt e Ali. “Artificial Intelligence and Human Psychology in Online Transaction Fraud” .

[6] Hamid Mehran, Anjan Thakor. “*Bank capital and value in the cross-section*”

[7] Elena Mattana, Ettore Panetti, *Journal of Banking and Finance (2014)*. “Bank liquidity, stock market participation, and economic growth”.

[8] Wei Wei, Yingli Liang, *Journal of Forecasting (2022)*. “ A Siamese network framework for bank intelligent Q&A prediction” .

[9] Leon Li, Robert Faff, *International Review of Economics and Finance (2019)*. “ Predicting corporate bankruptcy: What matters?”

[10] Hana Tamara Putri. “ANALISIS PENGARUH. RASIO KEUANGAN RBBR TERHADAP PERTUMBUHAN LABA BANK (STUDI KASUS PT. BANK CENTRAL ASIA (BCA), Tbk)”.

[11] Akhter T, Azad A.”Religiosity and bank earnings management: Revisiting international evidence”.

[12] Jones Jacob T, Sinclair Tara M, Stekler Herman O. “ A textual analysis of Bank of England growth forecasts“.

[13] Khan H. “ Bank competition, financial development and macroeconomic stability: Empirical evidence from emerging economies“.

[14] Morales P, Osorio D, Lemus J, Sarmiento M. “The internationalization of domestic banks and the credit channel of monetary policy”.

[15] Kim Y, Lee S, Lim H. “ Uncertainty, credit and investment: Evidence from firm-bank matched data”.

[16] Tsuchiya Y. “Assessing the World Bank’s growth forecasts”.

[17] Bluwstein K ,Buckmann M ,Joseph A ,Kapadia S , Şimşek Ö.”Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach”.

[18] Hasan M ,Hoque A, Le T. “ Big Data-Driven Banking Operations: Opportunities, Challenges, and Data Security Perspectives”.

[19] Keramati A, Ghaneei H, Mirmohammadi S. “Developing a prediction model for customer churn from electronic banking services using data mining”.

[20] Rahman M, Kumar V. ” Machine Learning Based Customer Churn Prediction in Banking”.

[21] Hu Y, Su J. “ Research on Credit Risk Evaluation of Commercial Banks Based on Artificial Neural Network Model”.

[22] Tran H, Le N, Nguyen V. “CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS”.

[23] Srivastava U, Gopalkrishnan S. “ Impact of big data analytics on banking sector: Learning for Indian Banks”.

[24] Appiahene P, Missah Y, Najim U. “ Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks”.

[25] Li B, Xie J. “Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network” .

[26] Kristóf T, Virág M. “EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks”.

[27] Sower J, Powell E, Hemeon K, Mann R, Pace S.” Ocean quahog (Arctica islandica) growth rate analyses of four populations from the Mid-Atlantic Bight and Georges Bank”.

[28] Chen Y, Wu J, Wu Z. “ China’s commercial bank stock price prediction using a novel K-means-LSTM hybrid approach”.

[29] Samad A, Armstrong V. “ Bootstrap-DEA management efficiency and early prediction of bank failure: Evidence from 2008-2009 U.S. bank failures”.

[30] Amrani O, Najab A, Azmi M. ”The impact of governance structure on bank performance: A cross-country panel analysis using statistical learning algorithms”.

[31] Domashova J, Gultiaev A. “ Predicting the revocation of a bank license using machine learning algorithms”.

[32] Levine O, Warusawitharana M. “ Finance and productivity growth: Firm-level evidence”.

[33] Arora D, Kashiramka S. “ What drives the growth of shadow banks? Evidence from emerging markets”.

[34] Carmona P, Climent F, Momparler A. “ Predicting failure in the U.S. banking sector: An extreme gradient boosting approach”.

[35] de Lima Lemos R, Silva T, Tabak B. “ Propension to customer churn in a financial institution: a machine learning approach”.

[36] Verma P. “Churn prediction for savings bank customers: A machine learning approach“.

[37] Maan J, Maan H. “ Customer Churn Prediction Model using Explainable Machine learning”.

[38] Kamble P, Nair A, Saini T, Patil G. “ Churn Prediction in Banking Sector”.