**Group Information**

**Topic / Dataset: Bank Customers churn prediction**

**Group No: 4**

**Table 1: Group information**

| **Student No** | **Student ID** | **Name** | **% of Contribution** |
| --- | --- | --- | --- |
| 1 | 2020-2-60-061 | Sujana Islam Smrity | 30% |
| 2 | 2021-1-60-071 | Mim Bin Hossain | 35% |
| 3 | 2021-1-60-059 | Abidur Rahman | 35% |

**Table 2: List of datasets (DO NOT DELETE ROW 0)**

| **No** | **Year: Dataset Short Name** | **Dataset Full Name** | **Dataset Link** | **Found by Student No** | **Comment** |
| --- | --- | --- | --- | --- | --- |
| 0 | 2023: ABC | Absolute Branch Dataset | [Link1](https://www.ewubd.edu/) | 3 | The dataset was downloaded and stored shared drive |
| 1 | 2023  BCC | Bank Customer Churn | [Link2](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) | 2 |  |
| 2 | 2022  CCCP | Credit Card Churn Prediction | [Link3](https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn) | 1 |  |
| 3 | 2022  BCP | Bank Churn Prediction (AUC = 0.99) | 97.1% Correct | [Link4](https://www.kaggle.com/code/dwipamujibagaskara/bank-churn-prediction-auc-0-99-97-1-correct) | 3 |  |
| 4 | 2023  BCP | Bank churn Prediction (Comprehensive EDA & model) | [Link5](https://www.kaggle.com/code/celocruz/bank-churn-prediction-comprehensive-eda-model) | 3 |  |
| 5 | 2023  BCA | Bank Churn analysis | [Link6](https://nbviewer.org/github/adin786/bank_churn/blob/main/churn_analysis.ipynb) | 3 |  |
| 6 | 2023  BCCA | Bank Customer Churn Analysis| SQL — Python | [Link7](https://medium.com/@amalia.wulandiari/bank-customer-churn-analysis-sql-python-fa6e2324c245) | 2 |  |
| 7 | 2022  BCEP | BankChurn EDA and Prediction Using Lazy Classifier | [Link8](https://www.kaggle.com/code/prathameshgadekar/bankchurn-eda-and-prediction-using-lazy-classifier) | 2 |  |
| 8 | 2023  MBD | Massive Bank dataset | [Link9](https://www.kaggle.com/datasets/ksabishek/massive-bank-dataset-1-million-rows) | 1 |  |
| 9 | 2023  BCCD | Bank Customer Churn Data | [Link10](https://www.kaggle.com/datasets/pentakrishnakishore/bank-customer-churn-data?fbclid=IwAR1YZLQzNiWE_DTBfcXR9tGIMq2iE7ZUNqjnwFPFHp6HLdcBE7Fex33LZtg) | 1 |  |
| 10 | 2020  FBTLS | Full Banking Transaction Log - sample | [Link11](https://www.kaggle.com/datasets/demodatauk/full-banking-transaction-log-sample/data) | 3 |  |
| 11 | 2022:  MLDCCCP | Machine Learning to Develop Credit Card Customer  Churn Prediction | [Link12](https://leaps.analyttica.com/) | 3 |  |
| 12 | 2020  CCBC | Churn Problem for Bank Customer | [Link13](https://www.kaggle.com/code/mathchi/churn-problem-for-bank-customer) | 1 |  |
| 13 | 2022  CCCP | Bank-Customer-Churn-Prediction | [Link14](https://github.com/FridahKimathi/Bank-Customer-Churn-Prediction/blob/main/Data/churn.csv) | 1 |  |
| 14 | 2023  RCPBML | Reduce customer churn in a bank using machine learning | [Link15](https://github.com/AbidurRahman1111/BankChurnModelling) | 3 |  |
| 15 | 2022  PCB | Predicting Churn Rate in a Bank | [Link16](https://github.com/samietex/Churn_Modelling_ANN/blob/main/ChurnRatePred.ipynb) | 2 |  |
| 16 | 2022  CCCP | Credit Card Churn Prediction | [Link17](https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn) | 3 |  |
| 17 | 2021  BCP | Bank Churn | [Link18](https://www.synthesized.io/data-template-pages/bank-churn-modelling) | 2 |  |
| 18 | 2021  PBC | predicting\_bank\_churn | [Link19](https://github.com/geete5h/Predicting-Bank-Customer-Attrition/tree/main/predicting_bank_churn_report_v1.1_files/figure-markdown_github) | 2 |  |
| 19 | 2020  BCCD | Bank-Customer-Churn-dataset | [Link20](https://github.com/hassanikram/Bank-Customer-Churn-dataset/blob/master/Churn_Modelling.csv) | 2 |  |
| 20 | 2018  BCCP | Bank Customer Churn Prediction | [Link21](https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/data) | 3 |  |
| 21 | 2020  UAPB | EDA 101: Univariate Analysis for Beginners | [Link22](https://www.kaggle.com/code/lonewolf95/eda-101-univariate-analysis-for-beginners) | 1 |  |
| 22 | 2021  ICPBA | Improved Churn Prediction Model In Banking Industry And Comparison Of Deep Learning Algorithms | [Link23](https://www.journal-aquaticscience.com/article_133719_d883304ae722b3c710a2722f1f0d3e7a.pdf) | 1 | In an article. |
| 23 | 2017  BTD | Bank Turnover Dataset | [Link24](https://github.com/AbidurRahman1111/Bankchurn/blob/main/Bankchurn_dataset%20(2).csv) | 2 |  |

**Table 3: List of articles that cited the datasets in previous table (DO NOT DELETE ROW 0)**

| No | Paper title | Journal / conference name | Published Year: Paper Link | Citation count | Paper Cited dataset No. x from Table 2 | Found by Student No |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | A brief history of time: an example paper name |  | 2023: [Link](https://www.ieee.org/) | 1206 | 0, 5 and 6 | 2 |
| 1 | Bank Customer Churn Prediction Using Machine Learning | MICHAEL WIRYASAPUTRA | 2022: [Link](https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/) | 20 | 1 | 2 |
| 2 | Predicting Customer Churn With Classification Modeling | Towards Data Science | 2020: [Link](https://towardsdatascience.com/customer-churn-analysis-eda-a688c8a166ed) | 41 | 7 | 2 |
| 3 | Churn Prediction for Savings Bank Customers: A Machine  Learning Approach | Journal of Statistics Applications & Probability | 2019: [Link](https://www.naturalspublishing.com/files/published/yt9r868jnrv116.pdf) |  | 1 and 7 | 1 |
| 4 | Predicting Churn Rate in a Bank Using Artificial Neural Network with Keras | Towards AI | 2022: [Link](https://pub.towardsai.net/predicting-churn-rate-in-a-bank-using-artificial-neural-network-with-keras-3bdc81e74f47) | 71 | 6 | 2 |
| 5 | CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED  CLASSIFICATION MODELS | Interdisciplinary Journal of Information ,  Knowledge and Management | 2023: [Link](https://www.ijikm.org/Volume18/IJIKMv18p087-105Tran8783.pdf) | 5 | 2 | 3 |
| 6 | CHURN PREDICTION IN BANKING USING ML WITH ANN | INSTITUTE OF SCIENCE AND TECHNOLOGY | 2022: [Link](https://sist.sathyabama.ac.in/sist_naac/documents/1.3.4/1822-b.e-cse-batchno-354.pdf) |  | 1 | 2 |
| 7 | predicting\_bank\_churn\_report | Geetesh Lokhande  undo | 2021: [Link](https://github.com/geete5h/Predicting-Bank-Customer-Attrition/blob/main/predicting_bank_churn_report_v1.0.pdf) | 03 | 18 | 3 |
| 8 | Customer Churn Prediction using Decision Trees | Enjoy Algorithm | 2020: [Link](https://www.enjoyalgorithms.com/blog/customer-churn-prediction-using-ml) |  | 1,7 | 1 |
| 9 | Bank Customer Churn Prediction | Github | 2023: [Link](https://github.com/FridahKimathi/Bank-Customer-Churn-Prediction) | 02 | 1,7 and 10 | 1 |
| 10 | Bank Customer Churn Prediction Model | Medium | 2020: [Link](https://medium.com/@611noorsaeed/bank-customer-churn-prediction-model-f1a1bbfef745) | 30 | 13 | 1 |
| 11 | Reduce customer churn in a bank using machine learning | Neuraldesigner | 2023: [Link](https://www.neuraldesigner.com/learning/examples/bank-churn/) |  | 14 | 2 |
| 12 | Artificial Neural Network Bank Customer Churn Prediction Model | Towards AI | 2020: [Link](https://pub.towardsai.net/artificial-neural-network-bank-customer-churn-prediction-model-4cc3ee57b811) |  | 1 and 10 | 3 |
| 13 | Customer Churn Warning with Machine Learning | Springer link | 2018: [Link](https://link.springer.com/chapter/10.1007/978-3-030-03766-6_39) | 3 | 4 | 2 |
| 14 | Machine Learning to Develop Credit Card Customer  Churn Prediction | Theoretical and Applied Electronic Commerce Research | 2022: [Link](https://mdpi-res.com/d_attachment/jtaer/jtaer-17-00077/article_deploy/jtaer-17-00077.pdf?version=1668590848-17-00077.pdf) | 13 | 11 | 2 |
| 15 | Investigating customer churn in banking | ScienceDirect | 2023: [Link](https://www.sciencedirect.com/science/article/pii/S2666764923000401) |  |  | 2 |
| 16 | Bank Customer Churn Analysis| SQL — Python | Medium | 2023: [Link](https://medium.com/@amalia.wulandiari/bank-customer-churn-analysis-sql-python-fa6e2324c245) | 55 | 6 | 2 |
| 17 | A Comparative Study of Machine Learning Algorithms for Bank Customer Churn Prediction | International Journal of Scientific Research in Engineering and Management | 2023: [Link](https://ijsrem.com/download/a-comparative-study-of-machine-learning-algorithms-for-bank-customer-churn-prediction/) |  | 21 | 2 |
| 18 | A big data analytics model for customer churn prediction in the retiree segment | International Journal of InformationManagement | 2019: [Link](https://www.sciencedirect.com/science/article/abs/pii/S0268401218301518) | 115 |  | 3 |
| 19 | A Dynamic Classification Approach to Churn Prediction in Banking Industry | Dynamic Churn Prediction in Banking Industry | 2020: [Link](https://core.ac.uk/download/pdf/326836343.pdf) | 141 |  | 3 |
| 20 | Negative Correlation Learning for Customer Churn Prediction: A Comparison Study | Research and Development of Advanced Computing Technologies | 2015: [Link](https://www.hindawi.com/journals/tswj/2015/473283/) | 3321 | 3 | 3 |
| 21 | Prediction Of Churning Customers forBank andTelecom | Population Therapeutics & Clinical Pharmacology  undo | 2023: [Link](https://www.jptcp.com/index.php/jptcp/article/view/1887/1980) |  | 19 | 3 |
| 22 | Prediction of Customer Churn in Banking Industry | STAT 642-674 Final Project Sina E. Charandabi | 2020: [Link](https://arxiv.org/ftp/arxiv/papers/2301/2301.13099.pdf) | 23 | 23 | 2 |
| 23 | Customer churn prediction using improved balanced random forests | ScienceDirect | 2009: [Link](https://www.sciencedirect.com/science/article/abs/pii/S0957417408004326) | 224 | 13 | 2 |
| 25 | Improved Churn Prediction Model In Banking Industry And Comparison Of Deep Learning Algorithms | International Journal of Aquatic Science | 2021: [Link](https://www.journal-aquaticscience.com/article_133719_d883304ae722b3c710a2722f1f0d3e7a.pdf) | 65 | 22 | 1 |
| 26 | Bank Customer Churn Prediction | Rpubs | 2022: [Link](https://rpubs.com/leexinyang/bankcustomerchurnprediction) |  | 4 | 3 |
| 27 | Predictive Framework for Advanced Customer Churn  Prediction using Machine Learning | International Journal of Computer Applications (0975 – 8887) | 2023: [Link](https://www.researchgate.net/profile/Abhyarthana-Bisoyi/publication/371769793_Predictive_Framework_for_Advanced_Customer_Churn_Prediction_using_Machine_Learning/links/649479e48de7ed28ba4cab23/Predictive-Framework-for-Advanced-Customer-Churn-Prediction-using-Machine-Learning.pdf) | 34 | 12 | 3 |
| 28 | Bank Customer Churn Prediction Using Machine Learning | International Journal of Scientific Research in Computer Science, Engineering and Information Technology | 2022: [Link](https://d1wqtxts1xzle7.cloudfront.net/88421406/CSEIT228389-libre.pdf?1657472257=&response-content-disposition=inline%3B+filename%3DBank_Customer_Churn_Prediction_Using_Mac.pdf&Expires=1701611533&Signature=CSi3xQueks9rvS6eWewviFpyU9u1Ir5FYsV6yIvZX7xlqOgVJYAxYHaABMkjQlWuPYFunwgyfbySrbJ5SiDq-QU8RRYLRkS-8sUDWdlJ46pQ3tpujlk0Nsa46JIfi0AF7GiCid7E7r8F8NJStYYTveC-M1mSipQiDzhu9BVpV-L0dhsjKWSiqOvrsGK51CHbHZW8x8UFi-YrxaZAuOlaMObqCHp21fopElrxoIxhDpBWJ4-Q2uousIe-zJJJo0JUIr1hiYaOgzoGI4a0-sy~PWu7jMLRan8H37no2t1amoDBaweae3WnJ8BKYE0ipnDp2JxWtZ7olXQ88~4myJcEzg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA) | 63 | 15 | 1 |
| 29 | Improving transaction safety via anti-fraud protection based on blockchain | Taylor & Francis Online | 2023: [Link](https://www.tandfonline.com/doi/full/10.1080/09540091.2022.2163983) | 1129 | 14 | 2 |
| 30 | Bank Churn Prediction using popular classification algorithms | Medium | 2021: [Link](https://medium.datadriveninvestor.com/bank-churn-prediction-using-popular-classification-algorithms-143d72dfc70b) |  | 2 | 1 |
| 31 | Unlocking the code of Customer Churn: Predictive Strategies for Banking Success | Pinaki Sahu 0000pinaki1234.kv@gmail.com | 2022: [Link](https://www.researchgate.net/profile/Pinaki-Sahu/publication/375526341_Unlocking_the_code_of_Customer_Churn_Predictive_Strategies_for_Banking_Success/links/654e0738b1398a779d751c6d/Unlocking-the-code-of-Customer-Churn-Predictive-Strategies-for-Banking-Success.pdf) |  | 9 | 3 |
| 32 | Investigating customer churn in banking: A machine learning approach and visualization app for data science and management | In Press, Journal Pre-proof | 2023: [Link](https://www.sciencedirect.com/science/article/pii/S2666764923000401) | 87 | 20 | 2 |

**Introduction**

—------------------------------------------------------------------------------------------------------------------------------------

Instructions section: Do not delete this section. It will be updated and color coded time to time

Word count suggestion: 1000 or more

In this section, write in your own words:

1. General discussion about these datasets:
   1. Why they are important
   2. How to they help us
2. General characteristics of these types of datasets:
   1. Types of data
   2. Quantity of data
3. General application cases / area / domain
4. You can write anything else related to this

—------------------------------------------------------------------------------------------------------------------------------------

Write below:

**Introduction:**

Customer churn or attrition is the biggest issue of the banking industry. The profitability of banks totally depends on strong customer base and customer churn breaks the relationship with the bank which can lead the bank to down worth. Now it has become an alarming issue for the banks. If they could forecast customer attrition, it could be solved. Then the bank can be more profitable and economically strong. It is really hard to gain new customers than hold the old ones, so it is important to find the reasons for customer attrition. Customer churn affects the bank as a whole and spoils the whole customer base. Trust issues can be created among other customers. A bank can come to the road at a time. It should be predicted and solved before churn, otherwise banks will be affected and the economy of our country as a whole. Commercial banks try to hold customers because customers staying and banking success is correlated. At this time bank churn prediction can play a vital role in customer holding. Where customer loyalty is equal to a bank's financial health, this bank churn prediction model will help the bank to survive in the competitive field of banking. Every customer contributes to the bank’s earnings so it will prevent customer churn and reduce losses.

Importance of Bank Customer Churn Datasets:

Banks can reduce risks using these datasets because there is relevant data for predicting customer churn like transaction details, last date of transaction, number of transactions. These will help to detect possible risks for customer churn. Banks can plan a strategic solution using this dataset based on the client category as well as they can keep their valuable customers. Loss of a large number of customers can impact banks revenue, banks can optimize their revenue by putting some strategies to protect customers. In this case, the dataset can help. Banks can use cost effective plans to keep the customers that already have. Using the dataset, they can identify the reasons for customer churn and solve it instantly. A complete view of data can help to implement a strategy to reduce customer churn. Banks can improve their service by applying insight of the reasons why customers are leaving their bank and taking initiative steps.

How Bank Customer Churn Datasets Help:

A predictive model can be created using the dataset. There is historical data of the bank that describes the whole scenario of customer and bank’s relationship. An analytical overview of customers' relationship with the bank will help to predict what are the reasons for customer churn. This dataset will summarize the whole thing together. This predictive analytical model will also describe what customers want from the bank; what facilities attract them to stay with the bank. Banks can easily understand what action should be done to reduce customer churn. Banks can also find the reasons for dissatisfaction with the bank. The data is really useful and effective for the model. The banks can improve their services to satisfy customers. This dataset gives the bank feedback to learn from customers behavior and it leads the banks to better improvement. This can also help to create useful methods of modification.

Types of Data:

There is statistical data which are a customer's age, gender, income, occupation, and other characteristics used to determine are the specific populations more likely to leave the bank. There is also transactional Data which are Account activity, transaction details, withdrawal and deposit details that is used to identify transactional behavior and this will help to find specific patterns of customer churn. There are some attributes which represent customer service interactions where customer inquiries, complaints and contacts with customer service determine how customer service impacts customer attrition. There is information about product and service usage by customers Details on how particular services are used by the customers for evaluation of the relationship between customer churn and product usage. There is some data for sentiment analysis using review and customer feedback for observing customer satisfaction. There is binary label data which indicates whether a customer has churned or not to predict churn.

Quantity of Data:

On average the total number of customer records in the datasets is more than 10,000. Larger datasets offer more robust insights and are often preferred for training accurate predictive models. The length of time that the dataset contains historical data is 2018 to 2023. Extended time frames offer a more information thorough understanding of customer behavior and churn patterns over a high period of time. The datasets used here are most recent and we tried to use updated datasets. Although fine data like daily, monthly, transaction-level details can lead to larger datasets, it enables more accurate analysis. Many features allow for a more thorough understanding of customer behavior and potential churn factors. More quantitative data ensures more quality of data that ensures Accuracy, completeness, and consistency of the data. High-quality data is essential for building reliable predictive models.

General application cases:

If it can be predicted why and which customers are at a higher risk of churning it will be easier to improve facilities and hold the customers. It can be possible by Implementing targeted strategies for identified high-risk customers to prevent their churn. Understanding factors that lead to customer dissatisfaction about the bank and improving existing services or introducing new features that satisfies customer expectations. By marketing campaign, try to maximize the effectiveness of marketing efforts. Specific targeted customer segments with personalized campaigns to not only acquire new customers but also hold existing ones. Classify customers according to their probability and behavior of churn. Identify if there is an ideal solution and customizing marketing strategies. Simplify customer management and internal procedure by determining how operational improvements can raise customer service and engagement with the bank. It is also important to find fraud. Identify unusual behavior that might be a deception. Applying churn prediction models to a more comprehensive fraud detection system improves security protocols and it will help to provide a secure banking system to the customers. Customers who are likely to leave could be more credit risky, churn predictions can evaluate credit risk. Improving overall customer satisfaction by addressing issues by churn prediction model to enhance the overall customer experience and loyalty. Those clients have more probability to leave, a better strategic method should be implemented to minimize the probability to leave. Maintaining a competitive advantage in the marketplace by providing better services to the customer needs, businesses can stay above the competition by utilizing the churn prediction model as part of a larger strategy.

The implementation of bank customer churn prediction in banking sectors is beneficial for strategic decision-making, financial profitability and improving customer services. These factors lead to the development of the banks and increase their economic and financial condition and build a strong relationship between customers and the bank. The banks will be able to reduce churn and hold their loyal clients and improve their service more effectively.

**Review of datasets**

—------------------------------------------------------------------------------------------------------------------------------------

Instructions section: Do not delete this section. It will be updated and color coded time to time

In this section, write in your own words:

1. An introductory paragraph (500 words or more)
2. Make a separate section describing each dataset. Do not copy paste even a single line.
   1. Write characteristics of that dataset
   2. Write limitation of that dataset
3. Finally summarize all the dataset in a table.
4. You can write anything else related to this.

—------------------------------------------------------------------------------------------------------------------------------------

Write below:

Introductory para (>500 words):

Review of datasets

01) Dataset 1 (Bank Customer Churn):

Introduction:

Maintaining customers is very crucial in the highly competitive world of banking. Maintaining the current customers is much easier but gaining new customers is hard and expensive also. Identifying the reasons for customer churn gives the bank's ability to create some new features and focused initiatives for customer satisfaction in an effective way. Focusing on different factors that impact a customer's decision to leave the bank, important datasets can provide information about customer churn to analyze the whole thing.

Datasets Analysis:

Characteristics:

RowNumber, CustomerId, Surname are the columns that serve as identifiers which are not used in predicting customer churn models. Credit score is an important variable because it is linked to lower churn rates. Geography is also an important variable because location can influence customers to churn and impacts on the decision to leave the bank. Gender is valuable for targeted marketing strategies. Age is a strong predictor because older customers are more likely to be loyal. Tenure variable is effective for the model to evaluate the customers transaction duration. Longer tenure has a higher possibility to be higher loyalty and lower churn rates. Balance variable represents the information about the customers bank balance which indicates financial stability because higher account balances correlate with lower churn rates. NumOfProducts, HasCrCard, IsActiveMember these features provide insights about a customer's engagement with the bank. EstimatedSalary is indicating possibility to churn because lower salaries may be associated with higher churn rates.Exited attribute holds the information of the target variable that gives information about whether a customer left the bank or not. Complaint, Satisfaction Score, Card Type, Points Earned are the additional features which provide information about customer satisfaction and engagement with the bank.

Limitation:

RowNumber, CustomerId, surname these columns can’t be used in statistical calculations because they have no predictive value. Credit score is an important variable because it is linked to lower churn rates, but it has no other financial behaviors that may influence churn. Geography may simplify complex regional variations and cultural factors. Gender can influence customers because a specific gender population can prefer different choices. Different age groups have not so individual variations in behavior about a bank. Tenure does not account for various engagement levels. There may be short term financial history in the Balance variable and this short-term financial history might not be considered in account. NumOfProducts, HasCrCard, IsActiveMember these attributes may not summarize external factors influencing churn. EstimatedSalary may not be considered for individual financial priorities. The Excited variable lacks reasons behind the churn. Complaint, Satisfaction Score, Card Type, Points Earned provides limited information on the nature and complaints.

**Dataset Summary:**

| **Column** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| RowNumber, CustomerId, Surname | Identifiers, no predictive value | Lack of predictive value |
| CreditScore | Correlates with lower churn rates | Limited to financial behavior |
| Geography | Location influences behavior | Oversimplifies regional and cultural variations |
| Gender | Exploration of gender's role in churn | May oversimplify diverse factors influencing churn |
| Age | Older customers exhibit higher loyalty | Does not capture individual variations within age groups |
| Tenure | Longer tenure correlates with loyalty | Does not account for variations in engagement levels |
| Balance | Higher balances correlate with lower churn | May not consider short-term financial fluctuations |
| NumOfProducts, HasCrCard, IsActiveMember | Insights into customer engagement | May not capture external factors influencing churn |
| EstimatedSalary | Lower salaries may be associated with higher churn | May not account for individual financial priorities |
| Exited | Target variable indicating customer churn | Lacks context for reasons behind churn |
| Complain, Satisfaction Score, Card Type, Points Earned | Additional insights into satisfaction and engagement | Limited information on complaint nature and resolution |

In conclusion it can be said that understanding customer churn by analyzing a large set of factors. While certain attributes like credit score, age, and balance are strong predictors, the dataset has limitations, such as oversimplification about certain factors and the lack of information such as the absence of the reasons behind customer exits. A comprehensive approach can consider both demographic and behavioral aspects is crucial for developing an effective strategy in the competitive banking sector.

02) Dataset 3 (Bank Churn Prediction (AUC = 0.99) | 97.1% Correct):

Introduction:

This dataset provides an overview on customer behavior with the bank and all customer information of the bank, focusing on customer behavior to understand and predict churn. Each row represents a unique customer identification data by the CLIENTNUM. The dataset includes some demographic variables that provide product related features and transactional data. The goal is to analyze the provided factors to predict customers, this will allow the bank to implement effective strategies to prevent customer churn. Analyzing the characteristics of the dataset and its limitations, features is essential for deriving meaningful insights.

Characteristics:

The dataset consists of a large amount of data of the customers in the banking sector. This dataset will provide a clear insight about customer relationship and behavior with the bank. How the customers interacting with the bank clearly can be visualized using the dataset. This dataset includes important demographic information about the customers which are age, gender, marital status, degree of education. This information is useful because these things can affect the relationship with the bank. This dataset also consists of product and service-related data like the kind of card customers use and relational data that shows how long they are using these cards and how long they are connected with the bank. This dataset has also some additional data like transactional details, credit card information and customers activity measurement.

Limitations:

There are some limitations in the dataset. The attribute ‘Attition\_Flag’ is a binary column. So, there is only available information that the customer left the bank or not. But no information why they left the bank. There is a lack of detailed information in the dataset. There is no explanation about the cause and factors. There is also a lack of qualitative data such as clients testimonial or explanation of inactivity. There are also some Unknown values in marital status. These cannot give a clear insight that can make difficulties in predictive model analysis. If these limitations are considered and solved, the model can be improved and reliable.

**Dataset Summary:**

| **Column** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| CLIENTNUM | Unique identifier for each customer | Limited direct insights into customer behavior or attrition |
| Attrition\_Flag | Binary indicator of account closure | Lacks details on reasons for attrition, limiting analysis |
| Customer Demographics | Age, Gender, Dependent Count, Education, Marital Status, Income | I is unclear, "Unknown" in Marital Status introduces uncertainty |
| Product Information | Card Category | Limited information on product-related features |
| Relationship Metrics | Months on Book, Total Relationship Count | Does not capture qualitative aspects of the relationship |
| Inactive and Contact Metrics | Months Inactive 12 Mon, Contacts Count 12 Mon | Lack of context on reasons for inactivity or nature of contacts |
| Credit Card Metrics | Credit Limit, Total Revolving Bal, Avg Open To Buy, Avg Utilization Ratio | May not capture external financial behaviors or reasons behind credit utilization |
| Transaction Metrics | Total Amt Chng Q4 Q1, Total Trans Amt, Total Trans Ct, Total Ct Chng Q4 Q1 | Lacks detailed context on the nature of transactions or factors influencing changes |

Additional Considerations:

1. Remove missing values and clean the dataset to apply an appropriate strategy.

2. Find useful features to understand specific variables contribution in the prediction model

3. Different classifications should be considered in evaluation metrics to improve model performance.

4. Consider customer feedback and reviews because these can provide a qualitative insight about customer experience with the bank.

This dataset is a useful tool in the banking industry for predicting customer churn or attrition. Although it provides clear statements about customer experience and information. By understanding its characteristics and limitations it can help to boost the performance of the predictive model. The model can be improved using this which might include qualitative and external data. This will help in long term growth of the bank.

03) Dataset 4 (Bank churn Prediction (Comprehensive EDA & model)):-

Introduction:

The dataset under consideration is customer data from a bank, and the main goal is to predict and minimize customer churn by analyzing the reasons why customers leave the financial institution. In order to maintain a strong base of loyal customers, which is essential for the bank's long-term growth and profitability, customer churn must be prevented. According to the technical definition of the goal, the task is to classify data from the testing file using a predictive classification model that has been trained on the provided training data. The model emphasizes a balanced precision and recall while going to maximize the F1 (macroeconomics) score. The dataset offers a comprehensive picture of customer interactions with the bank through analyzing a variety of customer attributes, such as financial behaviors and demographic data.

Dataset Analysis:

Characteristics:

CLIENTNUM is a unique identifier for each customer. It is a primary key for the dataset. Attrition\_Flag indicates that the customer has churned or stayed. Customer demographics like Age, Gender, Education Level, Marital Status, Income Category provide clear statements of the customer's profile. Card Information stored in Card Category, Credit Limit, Total Revolving Balance, Avg Open to Buy attributes that holds detailed information related to the customers financial behavior. Relationship Metrics with the bank information are stored in Total Relationship Count, Months Inactive 12 Mon, Contacts Count 12 Mon reflects the customer's engagement and interactions with the bank. Transaction Metrics quantifies the financial transactions and changes over a specific time period. Avg Utilization Ratio provides relevant information about how credits are utilized by a customer.

Limitation:

CLIENTNUM is a unique identifier for the dataset, but it has no meaningful value that can be used in statistical analysis. It can only be useful for individual identification. It doesn’t contribute to the predictive model. Attrition\_Flag is an attribute with binary data. It provides status and mutually exclusive but no detailed information about the data. Demographic attributes like age, gender, educational level, marital status, income category all hold categorical data about customers which is useful for churn prediction models but there are some missing values, missing values can give miss information. It can miss hidden factors. Card Category, Credit Limit, Total Revolving Balance, Avg Open to Buy represents card information but it cannot capture external factors. Total Relationship Count, Months Inactive 12 Mon, Contacts Count 12 Mon is limited to quantitative analysis but less qualitative about customer satisfaction. Transaction information cannot represent the reasons behind transaction and economic facts. Avg Utilization Ratio cannot explain the financial priorities of individuals.

**Dataset Summary:**

| **Column** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| CLIENTNUM | Unique identifier for each customer | May not significantly contribute to predicting customer churn |
| Attrition\_Flag | Indicator of churn or retention | Lacks granularity on reasons for customer churn |
| Customer Demographics | Age, Gender, Education, Marital Status, Income | Limited demographic details may miss nuanced churn factors |
| Card Information | Card Category, Credit Limit, Total Revolving Balance | May not capture external factors or preferences influencing card usage |
| Relationship and Interaction | Total Relationship Count, Months Inactive, Contacts Count | Limited to quantitative metrics, lacks qualitative insights |
| Transaction Metrics | Total Amt Chng Q4 Q1, Total Trans Amt, Total Trans Ct, Total Ct Chng Q4 Q1 | May not capture context behind transaction behavior |
| Credit Utilization | Avg Utilization Ratio | May not fully capture individual financial priorities or reasons for credit utilization |

Combining existing features can create a new variable and can increase the power of prediction of the customer churn prediction model. Customer feedback and reviews should be incorporated, and this can give a qualitative reason for customers satisfaction and reasons for churn. This dataset provides useful prediction data for customer attrition, and it offers a high range customer attribute that can provide accurate information. But it also has missing values that can reduce accuracy. It can be solved by feature engineering and advanced analytical methods. This dataset will help the banking industry to be more proactive to maintain customers.

**04) Dataset 6 ( Bank Customer Churn Analysis| SQL — Python):-**

Introduction:

The dataset contains useful information about the bank's customers which will help to facilitate the analysis of factors influencing customer churn. The primary key is the CustomerId which uniquely identifies each customer record and forms the foundation for data exploration and predictive modeling. The dataset covers an extended range of customer attributes from demographic details like age and gender to financial indicators such as credit score, balance, and tenure. Additionally, the customer behavior metrics like the number of products, credit card status, and activity level provides a holistic view of the customer's relationship with the bank. Understanding the characteristics and limitations of this dataset is crucial for meaningful analysis and effective strategies to reduce customer churn.

Dataset Characteristics:

RowNumber represents a sequential order to the data entries. It is used for tracking the chronological sequence of customer records. CustomerId is a unique identifier for each customer. It acts as the primary key, because it holds unique value of customer in the datasets. CreditScore represents the credit information of the customer. It is a crucial financial indicator which influences the customer's eligibility for various banking services. Geography and Gender variables store the customer's location and gender information. These variables provide relevant information to understand regional and gender-based records. Age and Tenure is the customer's age and the duration of attachment with the bank. It is essential for profiling customer based on age which measures customer loyalty. Balance and Num Products indicates the customer's account balance and the number of services they used. HasCrCard and IsActiveMember are binary indicators of credit card and customer activity. It describes customer behavior and engagement with the bank which lead to churn. EstimatedSalary reflects the customer's estimated salary. It provides insights into the customer's financial capacity and potential spending behavior. Exited is a binary indicator of whether the customer left the bank. It is the primary target variable for predicting and understanding customer churn. Complain and Satisfaction Score Indicate customer complaints and satisfaction scores. It offers insights into customer feedback and sentiment, contributing to churn analysis. Card Type and Points Earned describe the type of card held by the customer and points earned through credit card usage. It describes customer loyalty and engagement with the bank providing additional information to churn prediction.

Dataset Limitations

The dataset contains lack of contextual information about specific customer interactions or external factors about customer churn which reduce the depth of analysis. The satisfaction score is a valuable attribute but its individuality and multiple variances in interpretation may affect the accuracy of churn prediction. Complain is a binary variable so it may not describe the reasons why customers complain of the issues. The predictive model can be biased towards the majority class if there is an imbalance between the number of customers who exited and those who stayed.

**Dataset Summary:**

| **Column** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| RowNumber | Represents the record number | Essential for tracking sequence but may lack substantive value |
| CustomerId | Unique identifier for each customer | Primary key for individual customer tracking |
| CreditScore | Reflects predicted credit behavior | Crucial financial indicator influencing eligibility |
| Geography and Gender | Capture location and gender information | Demographic variables providing regional and gender context |
| Age and Tenure | Provide insights into age and relationship duration | Essential for profiling and measuring customer loyalty |
| Balance and NumOfProducts | Indicate account balance and number of products | Financial indicators contributing to understanding profitability |
| HasCrCard and IsActiveMember | Binary indicators of credit card possession and activity | Influence customer behavior and engagement |
| EstimatedSalary | Reflects customer's estimated salary | Provides insights into financial capacity and spending behavior |
| Exited | Binary indicator of customer churn | Primary target variable for predicting and understanding churn |
| Complain and Satisfaction Score | Indicate customer complaints and satisfaction scores | Offer insights into customer feedback and sentiment |
| Card Type and Points Earned | Describe type of card held and points earned | Influence customer loyalty and provide additional context |

Overall, the dataset provides a large amount of information for churn prediction and customer behavior analysis in the banking industry. Although this information can be used to build a useful prediction model that will forecast bank churn earlier. The dataset includes a wide range of information. By resolving its limitations is essential to creating an accurate predictive model to successful churn reduction. When the richness of the dataset is carefully considered it can lead to improve the bank's long-term strategies to maintain customers.

05) Dataset 9 (Bank Customer Churn Data):

Introduction:

This dataset contains over 28,000 customer information so it will give accurate prediction information. Large dataset provides more accuracy. Customer\_Id variable is assigned to the different customers to track each customer information in the dataset easily. Because it acts as a primary key which contains unique value. This dataset contains high range data of different characteristics including financial and transactional indicators like Current balance and transactional history of each customer. This data also provides demographical data. Using the dataset an effectful and useful prediction model can be easily created with the binary columns. Large size of the dataset provides an appropriate factor about customer attrition. The dataset includes past history and present datas about customers. It is ready to examine improved customer churn.

Dataset Characteristics:

Customer demographic information like age, gender, dependents, occupation, city provides the customer's personal and professional details. The financial indicated columns may provide a complete evaluation of the customer's transactional patterns. Different kind of financial indicators are provided by these columns like current\_balance, previous\_month\_end\_balance, average\_monthly\_balance\_prevQ, average\_monthly\_balance\_prevQ2, current\_month\_credit, previous\_month\_credit, current\_month\_debit,previous\_month\_debit, current\_month\_balance and previous\_month\_balance. Relationship duration presents the duration of the customer's relationship with the bank that indicates loyalty. Branch Information columns which are customer\_nw\_category, branch\_code provides the customer's net worth category and the specific branch correlated with their account. Churn Indicator used as the target variable that indicates whether the customer has churned or not. Last transaction attribute captures the time of the customer's last transaction that helps analysis of churn behavior.

Limitations:

In the dataset, last\_transaction attribute is used to measure tenure from last transaction to present but lack of a continuous time series may limit the exploration over time. The occupation attribute may lack specific value as financial behaviors which impacts on predictive accuracy. The dependents attribute cannot be differentiated between financial dependents and non-financial dependents. branch\_code identifies the branch associated with the customer and it is unique but may vary which impacts the accuracy of branch-specific analyses.

**Dataset Summary:**

| **Category** | **Attributes** |
| --- | --- |
| Customer Demographics | age, gender, dependents, occupation, city |
| Financial Metrics | current\_balance, previous\_month\_end\_balance, average\_monthly\_balance\_prevQ, average\_monthly\_balance\_prevQ2, current\_month\_credit, previous\_month\_credit, current\_month\_debit, previous\_month\_debit, current\_month\_balance, previous\_month\_balance |
| Relationship Duration | vintage |
| Net Worth and Branch Info | customer\_nw\_category, branch\_code |
| Churn Indicator | churn |
| Transaction Timestamp | last\_transaction |

In the banking industry, the churn prediction dataset offers a strong basis for researching and predicting customer attrition. Its many features provide a complete comprehension of how customers behave, enabling in-depth analysis. Even though there are some restrictions, implementing them in analysis and modeling can lead to improved customer retention strategies and more precise forecasts. Because of its size and wide range of features, the dataset is a valuable resource for gaining insights that can direct proactive steps to keep customers and support the financial institution's long-term growth.

**06) Dataset 2 (Credit Card Churn Prediction):-**

**Introduction:**

Introduction:In the competitive banking industry, keeping customers is crucial. Understanding why customers leave helps create focused strategies like loyalty programs for retention. This analysis aims to study customer churn factors using a dataset from a bank. The goal is predicting and reducing customer churn, vital for long-term growth and profit. To achieve this, a classification model trained on provided data will predict churn in the test file. This dataset has ten columns covering crucial customer details like attrition , age, gender, dependents, education, marital status, income, card type, and banking relationship duration. It's a goldmine for understanding customer behavior and predicting churn which helps banks make better decisions.

**Datasets Analysis:**

**Client Number:**

* Characteristics: An unused variable, likely serving as an identifier.
* Limitations: Limited utility in analysis; does not contribute to the understanding of customer behavior or churn prediction.

**Attrition\_Flag:**

* Characteristics: Key target variable for churn prediction models.
* Limitations: When there's much more of one kind of information than another, the computer might not guess so well. Additionally, if certain important data is missing, it might be challenging to figure out who might leave the bank.

**Customer\_Age:**

* Characteristics: Provides information about the age distribution of the customers.
* Limitations: Potential missing data could impact the accuracy of age-related analyses.The present age distribution of the population may not be reflected in current data.

**Gender;**

* Characteristics: The gender distribution of the customer base is shown.
* Limitations: Gender representation bias might affect accuracy of models. Lack of information can limit a complete knowledge of gender-related issues.

**Dependent\_Count:**

* Characteristics: The number of dependents linked with each customer account.
* Limitations: Due to a lack of information on dependency, the analysis of financial responsibility may be hampered.

**Education\_Level:**

* Characteristics: Provides information on the educational background of customers.
* Limitations: A lack of context on educational background could reduce model efficiency.

**Marital\_Status:**

* Characteristics: Provides information on a customer's marital status.
* Limitations: It could be more challenging to figure out how marital status influences banking behavior in the absence of complete information.

**Income\_Category:**

* Characteristics:Classifies clients based on their income levels.
* Limitations: Missing data might limit the depth of analysis across income brackets. Lack of context on income sources might impact model predictions.

**Card\_Category:**

* Characteristics: Represents the type of card held by customers.
* Limitations: No apparent limitations mentioned. Likely provides insights into customer card preferences.

**Months\_on\_Book:**

* Characteristics: Shows how long customers have been with the bank.
* Limitations: Old information might not help us understand how customers change. If it's not updated, it might not show us who stays loyal to the bank these days.

**Dataset Summary:**

| **Column** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| Client Number | Unused variable | Limited Variables |
| Attrition\_Flag | Target variable for churn prediction | Imbalanced Data, Limited Context |
| Customer\_Age | Provides age distribution among customers | Missing Data, Outdated Information |
| Gender | Distribution of gender within the customer base | Potential Bias |
| Dependent\_Count | Indicates number of dependents | - |
| Education\_Level | Educational background of customers | Missing Data, Limited Context |
| Marital\_Status | Marital status of customers | Missing Data, Limited Context |
| Income\_Category | Segregates customers based on income levels | Missing Data, Limited Context |
| Card\_Category | Type of card held by customers | - |
| Months\_on\_Book | Reflects customer loyalty | Static Information, Outdated Information |

**Conclusion:**

In conclusion,Understanding why customers leave a bank is vital for improving services and retaining them. Analyzing various data like age, income, and card preferences helps predict potential departures. However, imbalanced or incomplete data could hinder accurate predictions, and static or outdated information might limit grasping changing behaviors. Even though it's tricky, using this data is super important for banks. It helps them make smart choices, keep customers, and grow for a long time.

**07) Dataset 5 (Bank Churn analysis):-**

**Introduction:**

In the competitive world of banking, customer retention is key. To understand why customers might leave, banks gather a lot of data.Banks need to keep customers happy, so they don't leave. This study focuses on using a collection of information from a bank to predict which customers might leave next, aiming to prevent it. The dataset used in this study includes ten important details like whether customers are thinking of leaving, their age, gender, family size they support, education level, marital status, income, card type, and how long they've been with the bank. This data is a goldmine—it helps understand customer behavior and predict who might leave.

**Characteristics of the Dataset:**

This dataset presents a comprehensive view of bank customer profiles and their interactions with the financial bank. It includes essential demographic details such as unique customer identification numbers (CLIENTNUM), current customer status indicating whether they are existing churn, Age, gender, the count of dependents associated with customer accounts, education level, marital status, income category, card category, and the duration of the customer's relationship with the bank. Additionally, it delves into transactional specifics, encompassing contact counts within the last 12 months, credit limits, revolving balances, available credit for purchases, changes in transaction amounts over specific quarters, total transaction amounts, total number of transactions, changes in the number of transactions, and the utilization ratio. However, the dataset also contains an unidentified column labeled 'Unnamed: 21' that lacks clear information and context within the dataset.

**Limitations:**

There are some issues with this dataset." First, there's an unusual column called 'Unnamed: 21' that doesn't make sense and is difficult to understand.Additionally, information that is given insufficiently or improperly could make it difficult to identify a particular client action. Also, there may be errors or missing datasets in the data, which could lower the predicted accuracy. Moreover, this data may only cover a brief period of time, which means it may not reflect changes of the bank's customer base over time. Because of these worries, using this dataset requires awareness.

**Dataset Summary:**

| **Column** | **Description** | **Characteristics** | **Limitations** |
| --- | --- | --- | --- |
| CLIENTNUM | Customer ID numbers | Unique IDs | Unused variable |
| Attrition\_Flag | Customer status | Churn prediction input | No apparent limitations |
| Customer\_Age | Age of customers | Demographic analysis input | Static information; might not capture changes |
| Gender | Gender distribution | Demographic analysis input | No apparent limitations |
| Dependent\_count | Number of dependents | Analysis of family size | No apparent limitations |
| Education\_Level | Educational background | Analysis of educational level | Lack of detailed context |
| Marital\_Status | Marital status | Analysis of marital status | No apparent limitations |
| Income\_Category | Income level | Segmentation based on income | May lack nuanced income details |
| Card\_Category | Type of bank card held | Bank card categorization | No apparent limitations |
| Months\_on\_book | Customer relationship duration | Duration of relationships with the bank | Static information; might not reflect loyalty trends |
| Total\_Relationship\_Count | Total relationship count | Transactional relationships | Lack of detailed context; impacts prediction |
| Months\_Inactive\_12\_month | Months inactive in the last 12 months | Transactional activity | Lack of detailed context; affects prediction |
| Contacts\_Count\_12\_month | Count of contacts in the last 12 months | Customer engagement | Lack of detailed context; affects prediction |
| Credit\_Limit | Credit limit | Transactional information | Potential missing or undefined values |
| Total\_Revolving\_Balance | Total revolving balance | Transactional information | Potential missing or undefined values |
| Avg\_Open\_To\_Buy | Available credit for purchases | Transactional information | Potential missing or undefined values |
| Total\_Amt\_Chng\_Q4\_Q1 | Changes in transaction amounts | Transactional changes | Lack of detailed context; affects prediction |
| Total\_Trans\_Amount | Total transaction amount | Transactional information | Potential missing or undefined values |
| Total\_Trans\_Ct | Total transaction count | Transactional information | Potential missing or undefined values |
| Total\_Ct\_Chng\_Q4\_Q1 | Changes in transaction count | Transactional changes | Lack of detailed context; affects prediction |
| Avg\_Utilization\_Ratio | Average utilization ratio | Transactional information | Potential missing or undefined values |
| Unnamed: 21 | Unidentified column | Unknown | Ambiguity; lacks clear information |

**Conclusion:**

In conclusion, this dataset contains lots of information about bank clients and their transactions. It help in predicting when clients could decide to leave the bank. However, there are certain things that we are unsure about, such as this unclear 'Unnamed: 21' section. Furthermore, it is more difficult to figure out specific components because this column may be absent or unclear in some rows. Despite these problems, banks must utilize this information in order to retain clients and develop. Resolving these issues and thoroughly evaluating the data could help banks in making more informed decisions and maintaining client satisfaction.

**08) Dataset 7 (BankChurn EDA and Prediction Using Lazy Classifier):-**

**Introduction:**

The dataset contains a lot of information about bank customers, such as their age, gender, country, credit score, how long they have been with the bank, their account balance, what benefits they use and whether they have graduated. using a bank. Examining this data helps banks understand customer habits, financial trends and reasons why customers might leave. This helps banks make better choices to keep customers happy and satisfied.

**Characteristics of the Dataset:**

This dataset has lots of info about bank customers.

This dataset contains a wealth of information about bank customers. The Customer ID assists with maintaining a record of each customer. A credit score shows how effectively someone manages credit, which is significant when applying for loans. Customers' country of nationality is displayed. Gender indicates whether they are male or female, however it may not encompass all identities. Age depicts several age groupings. Duration indicates how long they have worked for the bank. The balance represents how much money they have. The Products Number indicates how many bank products they use. Credit Card indicates whether or not they have the bank's card. Active Member indicates how frequently they use the bank. Estimated Salary estimates how much they make. Churn indicates how many clients leave a bank.This data set from Kaggle analyzes bank client behavior, which is useful for determining why people quit banks. It is free to use, making research affordable to everyone.

**Limitations:**

Although this dataset offers valuable information about a range of bank customers, it does have some limitations that should be noted. The idea of gender as a binary variable has an opportunity to exclude or incorrectly characterize non-binary or other gender identities by simplifying the vast variety of gender identities. Although indicators like age and tenure can provide valuable insights into demographics and loyalty, they are static data that may not reflect changing or evolving behaviors over time. Also, the dataset could be missing complete financial data, such as information about multiple sources of income or in-depth credit analyses. Furthermore, due to basic biases or missing data within the data set, analysis or models of prediction built with this data may be less accurate or dependable. When interpreting results or creating models based on these constraints, careful handling is essential.

**Dataset Summary:**

| **Variable** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| Customer ID | Unique ID for customer identification | - |
| Credit Score | Metric for creditworthiness assessment | Might not encompass all factors influencing creditworthiness |
| Country | Geographical location representation | May lack detailed geographic data |
| Gender | Provides demographic information | Oversimplification of gender identities |
| Age | Demographic segmentation | Static representation, may not capture changes |
| Tenure | Reflects customer loyalty | Represents a static period, might miss dynamics |
| Balance | Showcases account's financial status | Might not offer a comprehensive financial picture |
| Products Number | Indicates engagement level with bank | - |
| Credit Card | Ownership of bank-issued credit card | - |
| Active Member | Reflects customer's engagement | - |
| Estimated Salary | Approximation of customer's total income | May not capture all income sources |
| Churn | Reflects customer attrition | - |

Finally, this information on bank customer qualities is a useful resource for assessing customer behavior and developments in the banking business. It provides all aspects of variables, from demographic information to financial indicators, allowing for in-depth insights into client attributes and prospective churn rates. While the information is valuable, it has inherent limitations, such as simplistic gender representation and static variables that may not capture shifting habits. Researchers and analysts should use it with precaution in order to know its limitations while utilizing its numerous insights for informed decision-making and predictive modeling in the banking industry.

**09) Dataset 8 (Massive Bank dataset):-**

**Introduction:**

The dataset provided contains transactional information from REC-SSEC Bank, a government-backed banking institution that works over the Indian Peninsula. It contains many different kinds of transactional records, such as dates, domains reflecting the types of businesses involved, transaction locations, overall transaction values, and transaction counts. This dataset is a large survey collection that provides information into the bank's operating countryside by representing the wide range of transactions that occur across the country.

**Characteristics of the Dataset:**

The dataset contains a variety of transactional data, which provides great insights while also having significant limits. The Date section provides chronological information that aids in analyzing trends and feature exploration, but it requires additional information for a thorough understanding of transactional changes. Domain classification allows for industry-based classification, identifying important industries that influence transactions. The dataset is useful for analyzing how transactions increase over time and the various business types that are engaged. It also displays transaction trends in certain geographies. It does not, however, provide highly extensive information about individual industries or local locations. While financial data is valuable, it can be difficult to figure out without additional information, and while transaction counts assist in detecting odd activity, they do not provide a full understanding of how transactions occur. Despite the fact that the dataset is valuable. These limitations show that more information or context is required for a complete understanding of how transactions work in various regions of the bank.

**Limitations:**

The dataset has a lot of information, but it lacks key important details that would assist explain why transaction amounts change over time. It does not provide enough specifics about various industries, areas, or the reasons for transactions. Some sections, such as the money details and transaction counts, are difficult to grasp without additional information on what each transaction is for. Overall, these limitations indicate that we require more information or context to truly comprehend how the bank operates and why transactions occur in the manner they do.

**Dataset Summary Table:**

| **Dataset Section** | **Characteristics** | **Limitations** |
| --- | --- | --- |
| Date | Provides chronological view for trend analysis | Lacks external contextual information for fluctuations |
| Domain | Enables industry-based segmentation | Lacks granularity for precise industry-level analysis |
| Location | Identifies regional transactional trends | Lacks depth for insights into specific localities or branches |
| Value | Offers monetary perspective for financial assessment | Faces interpretation challenges without contextual information |
| Count of Transactions | Quantifies transaction volumes, aids anomaly detection | Limited insight into transactional nuances, requires qualitative analysis for depth |

The REC-SSEC Bank dataset provides information on various transaction methods, their timings, the sorts of businesses involved, trends across areas, money amounts, and transaction frequency. However, there are certain restrictions. It fails to explain why transactions differ, lacks extensive specifics for in-depth examination, and some financial details require further context for a greater comprehension. Altogether, these issues tell us we need extra data or more info to really understand how the bank works. Fixing these problems will help us make better decisions and run the bank better.

**Review of techniques applied on these datasets**

—------------------------------------------------------------------------------------------------------------------------------------

Instructions section: Do not delete this section. It will be updated and color coded time to time

In this section, write in your own words:

1. An introductory paragraph (500 words or more)
2. Find different techniques and categorically explain them. For example, deep learning based techniques, traditional ML based algorithms, feature fusion etc.
3. Make a table to summarize them
4. You can write anything else related to this.

—------------------------------------------------------------------------------------------------------------------------------------

Write below:

**Introduction:**

Customer churn has appeared as a critical concern for banks in the competitive banking industry. It has been more difficult to maintain loyal customers. More advanced machine learning techniques should be used to create a model that can forecast customer churn so that banks can take initiative to stop or reduce them. Today's world of technology changes day by day so everyday new techniques and algorithms are generated for solving real life problems. To create an advanced and modern machine learning model we have implemented more effective and suitable machine learning techniques. In this predictive model mainly used algorithms or techniques are- Support vector machine, Logistic regression, Extreme gradient boosting or XGBoost, Gradient boosting, Decision tree and Random forest. Each technique has its own characteristics and approaches to solve complex problems. These techniques have some limitations too.

Support Vector Machine or SVM is a popular supervised learning algorithm which found utility in the bank churn prediction model. SVM identifies an optimal hyperplane that separates in different classes. It is highly effective for classification tasks. SVM can also handle both linear and non-linear relationships in the dataset. It has many benefits and also limitations. The time complexity of SVM on large datasets is very high. Logistic Regression is a classical statistical method that is very simple and interpretable. It is a popular choice for bank churn prediction. It can handle binary classification problems with the nature of churn prediction where outcomes are binary that an individual customer churned or not. But it may struggle with complex relationships and be sensitive to outliers. XGBoost and Gradient Boosting are quite similar by name and also nature. Both ensemble data and provide powerful solutions for bank churn prediction. Both techniques construct predictive models iteratively by combining multiple weak learners, such as decision trees. XGBoost, known for its scalability and efficiency. These techniques ensure high predictive accuracy in a robust approach within customer data. Decision Tree and Random Forest are tree-based models that provide effective solutions for bank churn prediction. Decision Trees create hierarchical structures of decisions and Random Forest combines multiple decision trees for increasing accuracy. These models make the data valuable for the customer churn prediction model.

In this Bank customer prediction model, we used advanced and effective and useful techniques which can increase its accuracy. The characteristics, utilizations, benefits, and limitations of each technique are described here. They should consider when they applied in the machine learning model to bank churn datasets. Using these methodologies, banks can enrich their analytical approaches to build accurate and efficient churn prediction models..

**Technique 1 (Support Vector Machine):**

Support Vector Machine (SVM) is a popular Machine learning algorithm which is one of the most popular supervised learning algorithms. It is used for classification and regression problems but generally used for classification problems in machine learning. SVM used to create the best decision that can separate n-dimensional space into class. A comprehensive machine learning algorithm can be developed using this algorithm. It classifies the data and using the data it will help to make decisions based on the data. It creates a hyperplane by choosing extreme points. The extreme points are the support vectors. There are two types of support vector machine (SVM). One is the Linear support vector machine and another is a non-linear support vector machine. SVM is used in predictive models because it can interpret data for classification which can increase the prediction accuracy.

Here, In the Bank churn prediction model, SVM is used to identify complex data and maintain signifying the possible churn. SVM helps to fragment the customers by their financial conduction for individual churn prediction for different customers. SVM analyzes the past data so in this model SVM analyzes the dataset of customer churn where all the information about customers in detail is present and facilitates strategies to identify customers who are likely to churn. That’s how SVM can predict early. SVM also determines the important data or features from the dataset and uses it for feature engineering for logical and efficient model training.

Benefits of using SVM in the Bank churn prediction model:

SVM is mainly used in the bank churn model to increase accuracy of classification customer data to make them predictable. SVM manages both linear and non-linear relationships in the dataset to ensure workability in complex relationships with customer dataset. SVM improves on possibilities of large numbers of appearances to make it capable of analyzing the dataset of customer churn.

Limitations of SVM:

The model can be challenging in some scenarios if there is limited churn data. Because SVM is totally dependent on the quality of the dataset as well as dependent on quantity of the dataset. SVM’s time complexity is normally very high and when it is implemented on a large dataset it becomes higher that can hamper application.

Summary Table:

| Aspect | Description |
| --- | --- |
| Structure | SVMs operate by identifying an optimal hyperplane to separate different classes, making them effective for classification tasks. |
| Utilizations | SVMs are applied in pattern recognition, customer segmentation, early churn identification, and automated feature extraction in the context of bank churn prediction. |
| Benefits | High accuracy, handling linear and non-linear relationships, effectiveness in high-dimensional spaces, and robustness against overfitting are key advantages of SVMs. |
| Limitations | Data sensitivity, computational time complexity which may impact on SVMs' application in some scenarios mainly with limited data. |

SVM presents significant approaches to the field of bank churn prediction. It offers a strong system for accurate modeling. There are also some challenges but ongoing advantages and research application can achieve the way for even more effective approaches in predicting and mitigating customer churn in the banking sector.

**Technique 2 (XGBOOST):**

Extreme Gradient Boosting, or XGBoost for short, is a highly effective machine learning method. It is very good at handling large volumes of data and using complex Generalized Gradients Boosted Decision Trees (GBDT). It has been used in many different kinds of applications, such as classification, regression analysis, and numerical rank..

To understand XGBoost, consider some of the fundamental concepts on which it is based. Supervised machine learning is one of them. These days, we use intelligent algorithms to identify patterns in data through the use of labels (such as names or categories) and attributes (such as sizes or numbers). This optimizes the process by assisting the algorithm in predicting previously unknown information..

XGBoost also heavily depends on decision trees. These resemble a twenty-question game, except they are prediction-based. To get a good calculation, they offer yes-or-no questions and identify which ones are most important. Consider utilizing a decision tree to estimate the cost of a home depending on features like size and number of bedrooms.

All of these innovative concepts are combined in XGBoost, an incredibly intelligent and user-friendly library. It's similar to a secret weapon for people who wish to foresee the future wisely and accurately with little difficulty.

### **Benefits of Using XGBoost for Bank Churn Prediction Models:**

* **Accuracy and Predictive Power:**XGBoost does a great job at making extremely accurate forecasts. It continuously improves accuracy by using a large number of decision trees to identify possible churners.
* **Handling Complex Data:**Churn prediction entails an analysis of several customer data sets. Complicated feature sets can benefit from XGBoost's ability to handle a variety of data kinds, including numerical, categorical, and limited information.
* **Feature Importance Insights:** XGBoost offers perceptions into the significance of features. This is very helpful in churn prediction since it shows which characteristics or actions of the client are most likely to cause a churn.
* **Scalability and Speed:** Large datasets, which are frequently found in churn prediction scenarios, are an excellent match for XGBoost due to its efficiency and scalability. Without losing performance, it can effectively manage a sizable number of features and observations.
* **Regularization and Overfitting Control:**By using regularization techniques like L1/L2 penalties, the method lowers the chance of overfitting. Preventing the model from recalling unimportant details from the training set and ensuring its compatibility with fresh data are critical tasks.
* **Handling Imbalanced Data:** Datasets for churn prediction frequently show imbalanced classes—more non-churners than churners. The capacity of XGBoost to manage unbalanced data guarantees more precise forecasts and guards against skewed results.

**Implementation of XGBoost for Churn Prediction:**

**Data Preparation:** Turn prediction relies heavily on feature engineering. Before feeding the data into XGBoost, engineers usually preprocess the data, extract important features, deal with values that are missing, and classify categorical variables.

**Model Training and Tuning:** Thorough modifying is required for XGBoost's hyperparameters to which include learning rate, tree depth, and regularization parameters. Using cross-validation techniques, the optimal set of hyperparameters can be identified to increase model accuracy.

**Evaluation and Validation:** Once the model has been trained, its effectiveness in predicting churn is assessed using various evaluation measures, including accuracy, precision, recall, F1-score, and ROC-AU

**Limitations and Considerations:**

**Complexity and Interpretability:** The ensemble aspect of XGBoost can make the model less interpretable. It can be difficult to understand the relationships between certain traits and predictions.

**Computational Resources:**Computing-intensive training of an XGBoost model is possible, particularly when using big datasets and substantial hyperparameter adjustment.

**Potential Overfitting:**Overfitting is a possibility despite regularization strategies, particularly if the complexity of the model is not appropriately controlled.

**Data Quality and Preprocessing:** It is essential to ensure high-quality data. Excessive or low-quality information can cause the model to perform badly.

Summary Table:

| **Aspect** | **Description** |
| --- | --- |
| Accuracy | High predictive accuracy from ensemble learning, which is necessary for accurate churner identification. |
| Handling Complex Data | Able to manage a variety of data formats (numerical, categorical, sparse), which is essential for a variety of feature sets. |
| Feature Importance Insights | Ability to handle a variety of data formats (sparse, categorical, and numerical), which is essential for a range of feature sets. |
| Scalability and Speed | Effectively manages substantial datasets without sacrificing speed, appropriate for instantaneous prediction. |
| Regularization and Overfitting | Includes overfitting management strategies to provide improved generalization to new data. |
| Handling Imbalanced Data | Effectively deals with imbalanced datasets, preventing biased predictions in churn identification. |
| Complexity and Interpretability | It can be more difficult to understand the relationships between many things when they're operating together. It might take particular expertise to figure out everything. |
| Computational Resources | Training can need a lot of resources, particularly when working with big datasets and substantial parameter tuning. |
| Potential Overfitting | To avoid overfitting even with normalization, thorough model complexity control is needed. |
| Data Quality and Preprocessing | Depends on high-quality data; model performance may be negatively impacted by low-quality data. |

In summary, even if XGBoost's accuracy and durability make it an effective tool for churn prediction, its use required careful analysis of the quality of the data, model tuning, and finding a balance between interpretability and complexity.

**Technique 3 (Decision Tree):**

Decision trees are effective tools for supervised learning because they can handle tasks related to classification as well as regression. They take on a tree-like hierarchical structure that includes essential nodes including a root node, branches, internal nodes, and leaf nodes. Essentially, a decision tree learns simple decision rules from the properties of the data itself in order to build a model that forecasts the value of a target variable.A decision tree's visual representation, which aids in understanding and makes it easier for an intuitive understanding to occur. People can see how the model predicts things according to its transparency. It makes sense in the logical steps that it takes.

Decision trees can handle an extensive range of data, including numbers and classifications, they are extremely helpful. compared to most other methods, this flexibility overcomes the need for considerable data preparation, which optimizes the modeling process.

.

**Benefits of Using Decision trees for Bank Churn Prediction Models:**

Decision trees are unique in the topic of machine learning because of their many advantages. First of all, they are quite straightforward and understandable. An understandable description of decision-making processes is provided by its visual representation, which resembles a tree structure. This makes it understandable to a broad audience, ranging from learners to specialists.One other notable benefit is the little requirement for data preprocessing. Decision trees are flexible and can handle a variety of data types without having a lot of processing effort, in opposition to some other methods that call for precise steps in data the pretreatment process, such as addressing missing values or normalization.Their versatility in handling various data formats, such as category and quantitative, adds even more allure.The reason decision trees are so great is that they can handle a wide range of problems without requiring complex data modifications..

Decision trees also provide accessibility; they are 'white box' models. Unlike more intricate 'black box' models, they clearly state the reasoning behind choices, providing an easy understanding of the model's operation.

Furthermore, these models are excellent for employing statistical tests to check whether or not their predictions are accurate. Decision trees demonstrate their usefulness in practical scenarios by remaining robust even in the presence of incomplete information rules.

Finally,decision trees are highly beneficial for machine learning. They are easy to understand, flexible, easy to comprehend, and highly skilled at managing a wide range of outcomes.

**Limitations of Decision trees :**Decision trees have several benefits, but they can have disadvantages. A major disadvantage is the tendency towards overfitting. A decision tree's capacity to generalize successfully to new, unidentified information can be limited by being excessively complex, which results in extremely detailed structures that fit the training data too closely. To address this, techniques like pruning—which means cutting down on or simplifying the tree or placing restrictions on the depth of the tree and the number of leaf nodes are used. Even with these safety measures, decision tree models are still susceptible to overfitting, which negatively affects how well the models function with fresh datasets.

**Summary Table:**

| **Benefits of Decision Trees** | **Limitations of Decision Trees** |
| --- | --- |
| Simple to understand and apply | easily stressed out, resulting in complex models that might not be very generalizable |
| Little to no data preparation needed | Sensitive to small variations in data, leading to different tree structures |
| Handles both numeric and categorical data | Difficulty in representing certain complex concepts |
| Transparent model - interpretable | Computational complexity in learning optimal trees |
| Validation with statistical tests possible | Creation of imbalanced trees with dominant classes |
| Robustness to assumptions that are varied forms |  |

In conclusion,decision trees are easily comprehensible and interpreted instruments in the field of machine learning. They can handle a wide variety of data, are very user-friendly, and will even take you through their decision-making process. However, they can overfit and become somewhat overly sensitive to slight variations in the data, which reduces their dependability. To address that, people employ editing and collaboration techniques. Despite all of this, decision trees remain highly useful for handling a wide range of categorization and mathematical problems due to their simplicity and logic.

**Technique 4 (Random Forest):**

### BeIn the field of machine learning, Random Forest is a flexible method that may be applied to issues involving both regression and classification. Based on the ideas of ensemble learning, this method employs several classifiers to address complex problems and provide strong answers. The Random Forest algorithm is essentially a collection of decision trees that are assembled into a "forest" and trained using bootstrap aggregation or bagging techniques. Bootstrap aggregating, or "bagging," is a type of ensemble meta-algorithm used to improve machine learning model accuracy.

### **Benefits of Random Forest:**

* **High Accuracy:** For both classification and regression applications, random forests typically provide predictions with a high degree of accuracy. The outputs from several decision trees are combined to lower the chance of overfitting and increase accuracy overall.
* **Robustness to Overfitting:** By combining predictions from multiple trees, they reduce overfitting, a significant problem in individual decision trees. This group method reduces overfitting to the training set while preserving accuracy.\
* **Data Versatility:** Without requiring significant data preprocessing, random forests can handle both numerical and categorical data. They may be used to solve a wide range of issues and datasets because of their adaptability.
* **Reduced Variance:** Random forests produce forecasts that are more stable and dependable by merging several decision trees that have been trained on various subsets of data.

**Limitations of Random Forest:**

**Random Forest's Drawbacks:** Interpretability and Complexity Although random forests have a high predictive value, understanding them might be difficult due to their complexity. Individual decision trees may be easier to understand than the combined forecasts from several trees.

**Resource Intensiveness:** When training a large number of decision trees or a huge dataset in a random forest, it might be computationally costly. This can require more time for training and greater processing power.

**Hyperparameter Tuning:** While random forests can be used with little configuration and are comparatively straightforward to use, some experimentation and tuning may be necessary to optimize hyperparameters (such as the number of trees or tree depth) for optimal performance.

**Summary Table:**

| Technique | Key Characteristics | Applications |
| --- | --- | --- |
| Random Forest | Ensemble learning method | Regression and classification problems |
|  | It comprises multiple decision trees | Complex problem-solving, diverse dataset applications |
|  | Trained through bagging/bootstrap aggregation | Improved accuracy, reduced overfitting |
|  | Prediction by averaging outputs from different trees | Robust predictions with minimal configuration |
|  | Mitigates limitations of decision tree algorithms | Handling large datasets requires increased accuracy without extensive tuning |

In simple terms, random forests are quite helpful in resolving complex data problems. They excel at generating precise forecasts, particularly when there is plenty of data to work with. These algorithms are highly dependable because they guard against errors brought about by an excessive concentration on certain aspects. However, they can occasionally be a little difficult to understand and take a lot to process extremely large amounts of data. They are still a very useful technique in machine learning. They are fairly adept at doing a wide range of tasks, striking a decent mix between precision and adaptability.

**Technique 5 (Ridge CLassifier)**

For banks to survive churn—the problems of clients leaving a bank—prediction is critical. In order to predict whether or not consumers may quit banks these days, the Ridge Classifier is very important. It is highly accurate in predicting when a consumer may become dissatisfied, which helps banks find ways to maintain their satisfaction. With its ability to predict who might go, this tool aids banks to develop plans. Without being highly specific or comprehensive, it analyzes how customers behave and looks for patterns. Maintaining customer satisfaction is of utmost importance to banks. In the ever-evolving world of banking, let's examine how this smart tool helps banks predict who might leave and how to avoid that.

### **Benefits of Ridge CLassifier:**

**Regularization for Robustness:**The model becomes more durable and more useful by executing large values by the use of Ridge regularization, which reduces overfitting. In the process, there is less chance of learning junk or unimportant features and more opportunity to identify the root causes in consumer behavior.

**Efficient Handling of Features:**The Ridge Classifier effectively handles high-dimensional information, which is very helpful when working with datasets that have a large number of features.By choosing the most important factors and avoiding the less important ones, it simplifies the process of determining if someone is likely to quit. In this sense, the model is not very complex.

**Balance Between Accuracy and Interpretability:**The Ridge Classifier is an excellent tool for determining the most significant attributes without adding unnecessary complexity. Understanding what causes clients to leave banks makes sensible choices to keep them satisfied.

**Scalability and Computational Efficiency:**Because of its linear structure, it can scale to real-time applications and handle massive datasets effectively. Especially useful in the fast-paced world of banking, it allows institutions to anticipate and promptly deal with potential customer attrition.

**Limitations of Ridge CLassifier:**

**Limited Non-linear Relationship Modeling:** The Ridge Classifier may not fully capture complex non-linear relationships found in consumer behavior because it requires linearity between parameters with the target variable. In situations when linkages are extremely non-linear, its predictive skills could be limited because of incomplete capture of complex interactions or non-linear patterns.

**Sensitivity to Feature Scaling:** The Ridge Classifier is susceptible to the feature scale, just like a lot of linear models. Without appropriate normalization or scaling, characteristics with significantly different scales or ratings may be given undue weight, which could affect the performance of the model.

**Inability to Handle Irrelevant Features:** Ridge regularization does not explicitly pick features, but it does assist in lessening the impact of less important features. It might have trouble with datasets that have a lot of noisy or unnecessary characteristics, which could have an impact on prediction accuracy.

**Assumption of Feature Independence:**In real-world situations, feature independence may not always remain true, as assumed by the Ridge Classifier. Because different variables in banking data may be connected, the assumption of independence may oversimplify the model and produce predictions that are skewed.

**Summary Table:**

| **Aspect** | **Description** |
| --- | --- |
| Advantages |  |
| Regularization for Robustness | Mitigates overfitting, ensuring a more generalized model. |
| Efficient Handling of Features | Handles high-dimensional data, discerns influential features, and reduces model complexity. |
| Balance Between Accuracy and Interpretability | Offers a balance between model complexity and interpretability, aiding in understanding feature importance. |
| Scalability and Computational Efficiency | Efficient handling of large datasets, suitable for real-time applications. |
| Baseline Model for Comparison | Provides a reference point for evaluating more complex models or feature engineering techniques. |
| Limitations |  |
| Limited Modeling of Non-linear Relationships | Assumes linearity and might struggle with non-linear patterns in data. |
| Sensitivity to Feature Scaling | Requires proper normalization as it's sensitive to feature scales. |
| Inability to Handle Irrelevant Features | Doesn't explicitly perform feature selection, impacting predictions with noisy features. |
| Assumption of Feature Independence | Assumes feature independence, which might not hold true in real-world scenarios. |
| Limited Performance with Imbalanced Data | Might struggle to learn from minority classes in highly imbalanced datasets. |
| Requirement for Hyperparameter Tuning | Tuning the regularization parameter demands computational resources and time. |
| Not Suitable for All Scenarios | Might not perform optimally with highly complex or non-linear datasets. |

In real-world situations, feature independence may not always remain true, as assumed by the Ridge Classifier. Because different variables in banking data may be connected, the assumption of independence may oversimplify the model and produce predictions that are skewed.

**Technique 6 (Logistic Regression):**

Logistic regression is a classification model that is widely used in statistical method binary classification problems. This is applicable in binary models like bank churn models where the output is binary that is churn or not. Logistic regression outcome is most commonly binary such as true or false, yes or no etc. Logistic regression is a classification algorithm which is a useful analysis method in classification problems. Logistic regression is a supervised learning algorithm used to classify categorical data with binary value. It can build a logistic model like bank customer churn prediction. It also can be used in class probability estimation. Logistic regression is one of the popular classification algorithms because of the simple assumption of linear decisions and non complex decisions, less prone to overfitting. When we classify training examples then often overfitting occurs.

Where linear regression provides continuous value in the output, logistic regression predicts that an individual data point belongs to a specific category. It translates or converts the categorical value into a binary value whose range is 0 to 1. Logistic Regression provides possible estimation value to banks to predict a customer churning. The eco efficiency of insights impacts on different features on customers who are thinking of churn. Logistic Regression makes it easier for stakeholders to understand the model's decision-making process. Logistic Regression is statistically suitable for real-time predictions.

Benefits of using Logistic Regression in Bank Churn Prediction:

Logistic Regression provides clear and interpretable results which are easy to understand of the elements convincing customers to churn predictions. Useful for binary classification like bank customer churn where predicting an individual customer churned or not. Logistic Regression not only classifies data but also provides possible evaluations. Logistic Regression is somewhat directly implemented by making it available.

Limitations:

It is possible that this linear relationship between the outcome and features is not always the case with logistic regression. Logistic Regression may try to capture complex patterns in the data when they already present a simple and linear relationship which is effective. Outliers in the dataset can impact a model's predictions.

Summary Table:

| Aspect | Description |
| --- | --- |
| Structure | Logistic Regression represents the probability of an event occurring using the logistic function to make it suitable for binary classification. |
| Utilizations | Logistic Regression is applied for probability estimation, an efficient interpretation and real-time predictions in bank churn prediction. |
| Benefits | Key benefits include interpretability, efficiency with binary outcomes, probability estimation and easy implementation. |
| Limitations | Challenges include the assumption of linearity, limited expressiveness for complex relationships, and susceptibility to outliers. |

Logistic regression works as an important method of bank churn prediction model. It converts all complex problems into simple ones, it interprets data to make it practical with a clear understanding. This can be benefitted often by using its simple features and assumptions. Logistic regression contributes to enhancing the accuracy of the model with its classification technology.

**Technique 7 (Gradient Boosting in Bank Churn Prediction):**

Gradient boosting is a machine learning technique that constructs decision trees. It is used in regression for solving classification problems. Since it is used in regression to classify data into an individual category, it can be used to build a predictive model like bank churn prediction. It contracts a group of decision trees and put it in sequential order. Each tree addresses the previous tree. It reduces loss by setting a new tree. The process continues repeatedly. The result is the total prediction of all trees. Gradient boosting can handle complex problems like regression.

Gradient Boosting has many uses in bank churn prediction. Gradient Boosting presents non-linear relationships in data to make it a more effective complex decision. The gradient boosting algorithm helps to increase features in the identification of the factors which contribute to the Churn prediction model. It combines multiple weak learners to create a reliable model that is also effective on general data. Gradient Boosting algorithms can manage imbalanced datasets in churn prediction where the number of churn may be lower.

Benefits of using Gradient Boosting in the Bank Churn Prediction model:

Gradient Boosting algorithms by organizing data can deliver high predictive accuracy compared to other models. The model describes the importance of different features and analyze the possible of churn. Gradient Boosting is suitable for complex problems and nonlinear relationships. Gradient Boosting is robust against overfitting because of its iterative nature.

Limitations:

Training in gradient boosting models can be computationally heavy with large datasets and deep trees. The performance of gradient boosting is highly dependent on the choice of parameters.

Summary Table:

| Aspect | Description |
| --- | --- |
| Structure and Operation | Gradient boosting builds an ensemble of decision trees sequentially to minimize the loss function at each iteration to create a powerful predictive model. |
| Utilizations | Gradient boosting is applied for handling non-linearity, feature importance determination, ensemble learning, and managing imbalanced datasets in bank churn prediction. |
| Benefits | High predictive accuracy, feature importance insights, effectiveness in handling non-linearity, and robustness against overfitting are key advantages of gradient boosting. |
| Limitations | Computational intensity during training and the need for careful parameter tuning are challenges associated with gradient boosting. |

Gradient boosting is a robust and effective technique in bank churn prediction. It is capable of handling complex relationships and handling imbalanced datasets. It contributes significantly to the accuracy and reliability of churn prediction models in the banking sector. The benefits of the gradient boosting are higher than its challenges of computational intensity. Gradient boosting is a valuable technique in customer churn model or any other predictive model

**Result and Discussions**

—------------------------------------------------------------------------------------------------------------------------------------

Instructions section: Do not delete this section. It will be updated and color coded time to time

In this section for each dataset you will :

1. Commence with a comprehensive paragraph, summarizing the results and presenting key findings from the dataset analyses and machine learning model performances. Elaborate on notable patterns, trends, or anomalies observed in the data.
2. Conclude the section with additional insights or observations derived from the analysis. Discuss any unexpected outcomes, challenges faced during the process, or areas where further exploration is recommended. This concluding part should provide a holistic understanding of the analysis, fostering a more in-depth comprehension of the results.

—------------------------------------------------------------------------------------------------------------------------------------

Write below:

[Dataset 1: Bank Customer Churn]

**Result and Discussion:**

The dataset consists of 10,000 rows and 18 columns. The dataset of 180,000 samples has been trained and tested using Support Vector Machine (SVM) and decision tree models. Dataset has been splitted for Training and testing. 30% of the dataset used for testing and 70% used for training. Cleaned and checked the dataset if there are any null or duplicate values in the dataset. Calculated the mean, median, standard deviation, min, max, first quartile and third quartile values for each row. Presented the information about all columns of the dataset and data type of each column. Calculated mean and median for each numerical column. The target column is the IsActiveMember which is a categorical column and is an integer. Showed a plot of null values and it is clear that there is no null value in the dataset. Also calculated the percentages of null values for each column and the percentage is 0 for all columns. Counted the number of customers for each Card Type.

Created subplots for the columns that stored numerical values. In those subplots, outliers are clearly visible. CreditScores and Age have many outliers but balance and point earned has no outliers. Generated a heatmap of correlation between numerical columns, ordinal columns and the column named Exited. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. Here each column is perfectly correlated with itself. Presented a pairplot of the whole dataset that visualizes the strongest correlations, linear relationships, non-linear relationships, outliers, groupings. Created three histograms of the target column, one where showing how many customers have a credit card and how many have no credit card of active members, one showing how many male and female customers are active and one showing how many active customers have exited and how many not exited. Two plots are created where one showing different age group customers numbers who are active and one showing the balance of each active member in the dataset.

Calculated the precision, recall, f1-score for the dataset using Support vector machine and decision tree model. The SVM model achieved the precision of 49% for inactive customers and 52% for active members, recall of 16% for inactive customers and 85% for active customers, F1 score of 24% for inactive members and 64% for active customers. The Decision tree model achieved the precision of 53% for inactive customers and 56% for active members, recall of 55% for inactive customers and 55% for active customers, F1 score of 54% for inactive members and 56% for active customers. Presented the class wise precision and recall values in two different plots for both SVM and decision tree model and compared the performance of the SVM and Decision Tree models and found that the SVM model performed better than decision tree. Our results suggest that the SVM model is more suitable for the research problem than the Decision Tree model as it provides better classification performance.

[Dataset 2: Credit Card Churn Prediction ]

**Result and Discussion:**

The dataset of bank churn prediction is utilized in the predictive model by training and testing. In this model, I used the dataset of 232,921 samples which is 10,127 rows and 23 columns. The dataset of 232,921 samples has been trained and tested using Support Vector Machine (SVM) and decision tree models. Dataset has been spitted for Training and testing. 30% of the dataset used for testing and 70% used for training and the random state of training is 42. The dataset is cleaned and checked if there are any null or duplicate values. Converted the column Attrition Flag from categorical to binary data. Examine how many customers are exiting and how many left. Calculated how many existing customers is male and how many are female. Calculated the number of customers for each education level and presented a plot to show the result. counting the members and showing the result in two subplots, one is for each income category, and one is for each marital status. Card Type is the target column, counting the number of customers for each type of card. Calculate the percentages of null values for each column and the percentage is 0 for all columns. Presented a pair plot of the whole dataset that visualizes the strongest correlations, linear relationships, non-linear relationships, outliers, groupings.

Calculated the unique values in each column and showed the data type. Generated a heatmap of correlation between all columns. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. Here each column is perfectly correlated with itself. But two other columns are perfectly correlated which are Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2 and attrition flag. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1 and attrition flag has a perfect negative correlation with each other. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1 is also having perfect negative correlation with Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2. Created boxplots for Customer\_Age that stored numerical values. In the boxplots, outliers are clearly visible. Created another boxplot for Total\_Revolving\_Bal which is also a numerical column shows the min, median, max and also outliers. Created four histograms of the target column, one were showing the number of cards for each card type of existing customers and the number of cards for each card type of attired customers, one showing how many male and female customers are existing and how many left. one showing the number of customers for each education level of existing customers and the number of customers for each education level of attired customers and another plot showing marital status of existing and attrited customers. Three plots are created to show different age group customers numbers, dependency and months on book for the members who are existing and for those who are left in the dataset.

Converted all the categorical columns to binary columns. Calculated the precision, recall, f1-score for the dataset using Support vector machine and decision tree model. The SVM model achieved a precision of 93% for customers who have blue cards and 0% for members who have gold cards, 0% for members who have silver cards and 0% for members who have platinum cards. The recall achieved by SVM is 100% for customers who have blue cards, 0% for customers who have gold cards, 0% for customers who have silver cards and 0% for customers who have platinum cards. The F1 score is 96% for customers who have blue cards, 0% for customers who have gold cards, 0% for customers who have silver cards and 0% for customers who have platinum cards. The Decision tree model achieved the precision of 99% for customers who have blue cards and 39% for members who have gold cards, 0% for members who have silver cards and 77% for members who have platinum cards. The recall achieved by the Decision tree model is 98% for customers who have blue cards, 57% for customers who have gold cards, 0% for customers who have silver cards and 73% for customers who have platinum cards. The F1 score is 99% for customers who have blue cards, 47% for customers who have gold cards, 0% for customers who have silver cards and 75% for customers who have platinum cards. Presented the class wise precision and recall values in two different plots for both SVM and decision tree model and compared the performance of the SVM and Decision Tree models and found that the decision tree model performed better than SVM. Our results suggest that the decision tree model is more suitable for the research problem than the SVM as it provides better classification performance for this dataset.

[Dataset 9: Bank Customer Churn Data]

**Result and Discussion:**

The dataset consists of 28,382 rows and 21 columns. The dataset of 596,022 samples has been trained and tested using Support Vector Machine (SVM) and decision tree models. Dataset has been splitted in two sections for Training and testing. From the whole dataset, 30% of the dataset was used for testing and 70% used for training. Dataset was cleaned by removing all null values and checked the dataset if there were any duplicate values, removing them. Calculated the mean, median, standard deviation, min, max, first quartile and third quartile values for each row. Presented the information about all columns of the dataset and data type of each column. Counted the number of male and female customers for both churned customers and existing customers and showed in a histogram plot. Also counted the number of customers in different occupations for both churned and existing customers. Generated individual plots for all numerical columns that show overview of those columns. These plots describe the characteristics of the columns.

Presented a pie plot of churn status. It is showing the insight of the column that 18.5% of the customers have churned and 81.5% of the customers have not churned. Created a pie plot that shows the occupation distribution where 61.7% customers are self employed, 23.7% customers are salaried,7.3% customers are students, 7.3% are retired and rest of the customers are companies. Generated a heatmap of correlation between all the columns. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 in the heatmap indicates no correlation and 1 in the heatmap indicates a perfect positive correlation. Here each column is perfectly correlated with itself. Generated a barplot counting churn of top 10 cities. The target column is the churn which is a categorical column and converted to an integer. Calculated unique values for each column. Generated another heatmap for numerical columns that clearly shows correlation between these numerical columns.

Calculated the precision, recall, f1-score for the dataset using Support vector machine and decision tree model. The SVM model achieved the precision of 87% for churned customers and 74% for existing customers, recall of 97% for churned customers and 37% for existing customers, F1 score of 92% for churned customers and 49% for existing customers. The Decision tree model achieved the precision of 88% for churned customers and 44% for existing customers, recall of 87% for churned customers and 45% for existing customers, F1 score of 87% for churned members and 44% for existing customers. Presented the class wise precision and recall values in two different plots for both SVM and decision tree model. Presented a pairplot of the whole dataset that visualizes the strongest correlations, linear relationships, non-linear relationships, outliers, groupings. Created three histograms of the target column, one where showing how many customers have a churned and how many are existing, one showing how many male and female customers are active and one showing how many active customers are from different cities. Two plots are created where one showing different age group customers numbers for churned and existing, one showing the branch code of each churned member in the dataset.

Compared the performance of the SVM and Decision Tree models and found that the SVM model performed better than the decision tree. Our results suggest that the SVM model is more suitable for the research problem than the Decision Tree model as it provides better classification performance.

[Dataset :13 Bank-Customer-Churn-Prediction]

**Result and Discussion:**

The bank churn dataset was subjected to training and testing procedures. It was made up of 14 rows and 1000 columns, with columns labeled "RowNumber," "CustomerId," "Surname," "CreditScore," "Geography," "Gender," "Age," "Tenure," "Balance," "Number Products," "HasCrCard," "IsActiveMember," "EstimatedSalary," and "Exited." This dataset was notable for having no null or duplicate values.The target column selected for analysis in the testing phase was "IsActiveMember," which was classified as a categorical target column. After classification, 5151 active members and 4849 inactive members were found in the dataset. This class breakdown was represented as a bar chart with no bars denoting null values.This analysis focused on two numerical columns: 'Age' and 'EstimatedSalary.' Box plot observations revealed a number of outliers in the 'Age' column but none in the 'EstimatedSalary' column. Understanding the distribution and possible abnormalities within these two columns required this information.Colors are used in a two-dimensional matrix to visually represent data in a grid heatmap. Each dimension denotes a particular trait category, and the color intensity shows the measurement magnitude obtained from these combined qualities. I determined the corresponding data types for each column by analyzing unique values in each one. The relationships between each column were represented by a correlation heatmap, which showed correlations ranging from perfect negative (-1) to perfect positive (1), with a value of 0 denoting no correlation. Furthermore, every column precisely.Every column also had a perfect correlation with every other column.Support Vector Machine (SVM) and decision tree methods were used to train and test a dataset of 232,921 samples for the predictive models. A predetermined random state of 42 was applied to the dataset, which was divided into 70% for training and 30% for testing. The goal was the 'IsActiveMember' column, which provided customer counts for various card kinds. 0% null values were found for each column after null value percentages were computed for all of the columns.

A pair plot was created to view groupings, correlations, outliers, and linear and non-linear relationships over the entire dataset. Furthermore, four histograms illustrating the distribution of the target column were generated. These displayed the quantity of active and inactive members, along with the gender distributions, credit card ownership, age-related activity trends, and corresponding balances for each.Binary columns were created from all of the category columns. For both decision tree and support vector machine (SVM) models, precision, recall, and F1-score were calculated. The SVM model produced 4% precision for inactive members, 20% for those who left the bank, and 10% for closed accounts, while it produced 93% accuracy for actively involved clients. With respect to card kinds, the SVM model produced 'blue' cards with an F1-score of 96%.

However, the decision tree model showed 39% accuracy for "gold," 77% accuracy for "platinum," and 99% accuracy for "blue" cardholders. 98% of the "blue," 57% of the "gold," and 73% of the "platinum" cards were remembered. At 99% F1-score, the decision tree model performed exceptionally well in predicting 'blue' cardholders; for 'gold' and 'platinum,' it performed moderately well; but, for 'silver' card holders, it achieved 0% F1-score.

For both SVM and decision tree models, visual representations of the class-wise precision and recall values were provided. The decision tree performed better than SVM, as demonstrated by comparative analysis, suggesting that it might be a better fit for this dataset's classification needs. The results clearly indicate that the decision tree model works better in accurately identifying customer behavior than SVM, which makes it more appropriate for this particular study.

**[**Dataset 5**:** Bank Turnover Dataset **]**

**Result and Discussion:**

There are eighteen columns and 10,000 rows in the dataset. SVM and decision tree models have been used for training and testing the dataset of 180,000 samples. A separate dataset has been created for testing and training. Of the dataset, 70% was utilized for training and 30% was used for testing. cleaned and examined the dataset to make sure it was free of null or duplicate values.For every row, the following values were calculated: mean, median, standard deviation, min, max, first quartile, and third quartile. provided details on every column in the dataset, including the type of data that is stored in each column.counted the number of male and female clients—both churned and current—and plotted the results on a histogram. Counted the quantity of existing and churned consumers in various occupations as well. created separate charts that display an overview of each numerical column. The columns' properties are depicted in these charts.To see groupings, correlations, outliers, and linear and non-linear relationships over the whole dataset, a pair plot was made. In addition, four histograms representing the target column's distribution were produced. These showed the number of users who were active and inactive, as well as the distributions of genders, credit card ownership, age-related activity trends, and associated balances for each.Every category column was converted into a binary column. F1-score, precision, and recall were computed for both decision tree and support vector machine (SVM) models. The SVM model yielded 93% accuracy for clients who were actively involved, but only achieved 4% precision for inactive members, 20% for those who left the bank, and 10% for closed accounts. The SVM model generated 'blue' cards with an F1-score of 96% in relation to card sorts.But the decision tree model indicated that for "gold," "platinum," and "blue" cardholders, the accuracy was 39%, 77%, and 99% respectively. Of the cards that were remembered, 98% of the "blue," 57% of the "gold," and 73% of the "platinum" were. The decision tree model did a remarkable job of predicting 'blue' cardholders at 99% F1-score; for 'gold' and 'platinum,' it did a fair job; but, for 'silver' cards, it achieved 0% F1-score.

Determine the dataset's precision, recall, and f1-score using the decision tree and support vector machine models. With respect to inactive customers, the SVM model's precision was 49% and 52% respectively, recall was 16% and 85% for inactive customers, and the F1 score was 24% and 64% for inactive members, respectively. Recall was 55% for inactive customers and 55% for active customers, F1 score was 54% for inactive members and 56% for active customers, and the decision tree model was able to accomplish these results. assessed the effectiveness of the SVM and Decision Tree models and displayed the class-wise precision and recall scores in two separate graphs for each model.that the SVM model is more suitable for the research problem than the Decision Tree model as it provides better classification performance.

**[**Dataset :14Reduce customer churn in a bank using machine learning**]**

**Result and Discussion:**

The name of the dataset is “Bank Churn.csv” and the dataset number is 05. The dataset of Bank churn prediction is utilized in the predictive model by training and testing. The dataset consists of 286 rows and 11 columns.This dataset consists of 21,20 samples, which have been trained and already tested using Support Vector Machine (SVM) and decision tree models. The dataset has been split for Training and testing. 30% of the dataset was used for testing and 70% was used for training.It is important to note that the test set should only be used for final evaluation. The model should not be adjusted based on its performance because of the test set to avoid information outflow. In some cases, a validation set should be used during the training data process to fine-tune the model without touching and the test set until the very end of the process. In this data set,the name of these index are Date', 'Domain', 'City', 'Balance', 'Total\_Transaction', 'Gender', 'Age', 'NumOfProducts', 'HasCrCard', 'EstimatedSalary', 'Exited.

In the column, this data set has no null values. This dataset has no duplicate records.This dataset consists of two (02) Dimensional. We have calculated the mean, median, standard deviation, min, max, first quartile (25%), and third quartile (75%) values for each row in the dataset.The target column is the Domain which is a categorical column.Now, we are drawing a bar plot , here, x-axis is for Age and y-axis is for Gender. We can see on the bar-plot that, for the maximum time, the Male is greater than the Female. The target column is the Domain which is a categorical column.There are many samples in the domain of 1 means INVESTMENTS (True) and 0 means PUBLIC (False).The INVESTMENT has 38 in the target column and PUBLIC has 37 in the target column.Now, we plot bar-chart and we can see from here that INVESTMENTS is greater than PUBLIC.

A pair plot, also known as a scatterplot matrix, is a visualization technique used in data analysis to test the relationships between pairs of variables. This dataset is usually used in machine learning because the measurements and classes provide a great way to determine classes. Because there are 8 measurements which create a 8x8 plot.

We take x and y which is test and train and test size is 03 and random\_state is 42.

Calculated the precision , recall, and f1-score for the dataset using the Support vector machine and decision tree model. The SVM model achieved a precision of 0% for INVESTMENTS and 0 % for PUBLIC, but we predicted that precision of 71% but it failed. A recall of 0% for INVESTMENTS and 0% for PUBLIC and we predicted the precision of 100% but it failed, an F1 score of 0% for INVESTMENTS and 0% for PUBLIC, but we predicted the precision of 83% but it also failed. The Decision tree model gained a precision of 17% for PUBLIC and17% for INVESTMENTS, recall of 25% for PUBLIC and 8% forINVESTMENTS, F1 score of20% for PUBLIC and 11% for INVESTMENTS. It shows the class-wise precision and recall values in two different plots for both the SVM and decision tree model and difference between the performance of the SVM andDecision Tree models and found that the SVM model performed better than the decision tree. Our results suggest that the SVM model is more flexible for the research problem than the Decision Tree model as it provides better classification performance.