# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Bank Customer Churn Prediction* |
| **Requestor** | *Professor Mustafa* |
| **Date of Request** | *May 08, 2024* |
| **Target Quarter for Delivery** | *Quarter 4* |
| **Epic Link(s)** | <https://www.kaggle.com/datasets/murilozangari/customer-churn-from-a-bank> |
| **Business Impact** | *Reduce the customer churn thereby improving business revenue and profits.* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

**Business Statement:**  
The primary objective is to address the issue of customer churn within the bank. Customer churn refers to the customers leaving the bank and taking their business elsewhere. By identifying the key factors that lead to customer churn, the bank can develop targeted retention strategies to maintain and enhance its customer base.

**Opportunity:**  
The dataset contains 10,000 rows and 14 columns and I aim to predict whether a customer will churn. The dataset has various demographic and behavioral variables. By analyzing the dataset, I can identify the patterns and trends that significantly impact a customer’s decision to leave the bank.

**Problem Statement:**  
The key problem is the lack of insight into the types of customers who are likely to churn and the reason of the decision to leave. Without this understanding, the bank cannot effectively address the root causes of churn or implement strategies to retain at-risk customers.

**Desired Outcome:**  
The desired outcome is accurate identification of customers who are at risk of churning through a predictive model. The insights gained from this analysis will enable the bank to create targeted retention strategies, thereby reducing the churn rate and increasing the customer loyalty and satisfaction.

**The specific ask:**

*Clearly articulate the specific task you will be conducting to help achieve the opportunity*

I will start by analyzing the demographics of customers and then understanding which customers are at the risk of churning the most, followed by the investigation of the factors involved.

To achieve the objective listed above, the following tasks must be conducted:

1. **Data Collection and Cleaning:** The dataset has already been imported from Kaggle. The first step will be to perform data cleaning to handle the missing values, outliers and inconsistencies to ensure the dataset is ready for analysis.
2. **Exploratory Data Analysis (EDA):** Conduct a thorough exploratory data analysis to understand the distribution and relationship within the data. This also includes visualization of key variables to identify trends and patterns that might influence customer churn.
3. **Feature Engineering:** Creation of new features that may provide additional insights into customer behavior and demographics. This also involves the transformation of existing features to enhance their predictive power.
4. **Model Development:** This step includes splitting the dataset into training and testing, developing multiple machine learning models (logistic regression, decision trees, random forests, gradient boosting) to predict customer churn followed by the evaluation of the performance of these models using appropriate metrics like accuracy, precision, recall and F1 score.
5. **Model Evaluation and Selection:** On the basis of the performance of different models, I will select the best performing model based on evaluation metrics and cross-validation results.
6. **Insights:** In order to identify the most significant factors contributing to customer churn, I will interpret the model results and generate actionable insights that can inform the bank’s customer retention strategies.
7. **Communication of Findings:** Prepare a comprehensive report and presentation that details the analysis process, findings, and recommendations followed by communicating the results to stakeholders to facilitate decision-making.
8. **Proposition:** Propose specific customer retention strategies based on the insights gained from the analysis and collaborate with relevant departments to implement these strategies.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

In the highly competitive banking industry, customer churn is a critical issue as it can impact bank’s market share and profitability. Various trends and research findings highlight the importance of understanding customer behavior to develop the effective retention strategies.

**Industry Trends:**  
1. **Digital Transformation:** Digital banking has increased customer expectations for convenience and seamless experiences. Banks that fail to deliver high-quality digital services are more likely to see higher churn rates.

2. **Personalization:** Customers expect personalized services tailored to their specific needs and preferences. Banks leveraging data analytics to offer customized solution can improve customer satisfaction.

3. **Customer Experience:** Positive customer experiences are crucial in retaining customers. Factors such as responsive customer service, user-friendly interfaces, and efficient problem resolution plays a significant role in customer loyalty.

**Research Findings:**

1. **Demographic Factors:** Factors such as age, income, and marital status can influence customer churn. For example: younger customers are generally more prone to switch banks compared to older customers.
2. **Behavioral Factors:** Transaction patterns, frequency of interactions and usage of bank services are key behavioral indicators of potential churn. Identification of these patterns helps in predicting and preventing churn.
3. **Competitive Landscape**: Analysis of competitors reveals that banks offering innovative products, competitive interest rates, and superior digital experiences have lower churn rate.

Key Competitors:

1. Chase Bank:

**Key Message:** Emphasis on convenience, comprehensive financial services, and robust digital banking platforms

**Product and Services:** Checking and savings account, loans, credit cards, and investment services.

**Market Share:** Significant national market share with strong regional presence across the United States.

1. Bank of America:

**Key Message:** Focus on customer satisfaction, innovative solutions and community management.

**Product and Services:** Personal and business banking, wealth management, and online banking services.

**Market share:** One of the largest bank in the U.S. , with a broad market reach and high customer base.

1. Wells Fargo:

**Key Message:** Commitment to trust, transparency and financial well-being of customers.

**Products and Services:** Comprehensive products and services including mortgages, personal banking, and commercial banking.

**Market Share:** Strong national presence with significant influence in various regions.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

**Revenue Gains:** Retention of customers directly translate to higher revenue as it is more cost-effective to retain existing customers than to acquire new ones. Satisfied customers engage in cross-selling opportunities further boosting revenue.

**Quality Improvement:** The project will enhance the overall quality of customer service by identifying the pain points and addressing it proactively.

**Cost and Time savings:** Reduction of customer churn can lead to significant cost savings by decreasing the need for aggressive marketing campaigns and aimed at acquiring new customers. It reduces the operation costs associated with handling customer complaints and issues.

**Customer Care:** Customers will benefit from a more personalized banking experience tailored to their specific needs and preferences. The increased personalization to their specific needs and preferences. Happy customers are less likely to churn and more likely to recommend the bank to others.

**Implication if we do nothing:**  
If no action is taken to address customer churn, the bank may face increased customer attrition, leading to loss in revenue. The cost of acquiring new customers to replace those who have churned will be higher than retaining existing ones. The bank’s reputation can suffer as high churn rates can signal poor customer service.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

Key Questions:

1. What are the primary demographic and behavioral factors that contribute to customer churn in the bank?
2. How can we accurately predict which customers are most likely to churn?
3. What patterns or trends can be identified from the customer data that indicate a higher likelihood of churn?
4. How effective are the current customer retention strategies, and what improvements can be made based on the data analysis?
5. What specific actions can be taken to retain the customers identified as high risk for churn?

Assumptions:

1. Customers with lower engagement levels (e.g. fewer transactions, infrequent use of banking services) are more likely to churn.
2. Demographic factors such as age, income, and marital status significantly influence the likelihood of customer churn.
3. High customer satisfaction and personalized service offerings reduce the likelihood of churn.
4. Customers who have recently experienced issues with customer service or product quality are more likely to leave the bank.
5. Competitive banks also play a significant role in the customer churn.

Hypothesis:

1. Customers aged 18-35 are more likely to churn compared to older customers due to their preference for digital-first banking solutions and higher expectations for service quality.

Null hypothesis: There is no significant difference in churn rates between customers aged 18-35 and older customers.

1. Customers with high income levels are less likely to churn due to their greater financial stability and investment in long-term banking relationships.

Null hypothesis: Income level does not significantly impact the likelihood of customer churn.

1. Customers who frequently interact with bank services (e.g., making transactions, using online banking) are less likely to churn.

Null hypothesis: The frequency of interaction with bank services does not significantly impact the likelihood of customer churn.

1. Customers who have recently have negative experiences with customer service are more likely to churn.

Null hypothesis: Recent customer service experiences do not significantly impact the likelihood of customer churn.

1. Implementing personalized retention strategies based on predictive analytics will reduce the overall churn rate by at least 15%.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

1. It is assumed that the data collected from Kaggle dataset is accurate, up-to-date, and reflective of the current customer base. Any discrepancies or outdated information could skew the analysis results.
2. Customer behavior patterns identified in the historical data will remain consistent in the future.
3. The analysis assumes a relatively stable economic environment.
4. It is assumed that the actions of competitors remain relatively constant.
5. Major changes in the competitor strategies could impact customer churn in ways not accounted for in current analysis.
6. It is assumed that all necessary data is available and accessible for analysis. Any gaps or limitations in data availability could hinder the comprehensiveness of the analysis.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | In this project, success is defined as the prediction of customer churn accurately and effective implementation of retention strategies that lead to a measurable decrease in the churn rate. The project should provide actionable insights that lead to the improvement in the customer satisfaction and overall business performance.  **Key Performance Indicators:**   1. **Churn Rate Reduction:** It is defined as the percentage decrease in the customer churn rate after implementing the predictive model and retention strategies. The target is to achieve at least a 15% reduction in the churn rate within six months of implementation. 2. **Model Accuracy:** It is defined as the accuracy of the predictive model in identifying customers at risk of churning. The target is to achieve a model accuracy of 85% or higher. 3. **Precision and Recall:** Precision is defined as the percentage of true positive churn predictions among all positive predictions, and recall is defined as the percentage of true positive churn predictions among all actual churn cases. The target is to achieve precision and recall rates of at least 80%. 4. **Customer Retention Rate:** This is defined as the percentage of retained customers after implementing retention strategies. The target is to increase the customer retention rate by 10% within one year. 5. **Customer Satisfaction and loyalty:** This is the measure of customer satisfaction based on survey responses. The target is to improve by 10% within a year. 6. **Cost savings:** It is defined as the reduction in costs associated with customer acquisition and churn management. The target is to achieve the cost savings of at least 20% related to customer retention efforts. 7. **Revenue Growth:** Increase in revenue is related to the improved customer retention. The target must be to achieve a 5% increase in revenue within one year. |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** *logistic regression, linear regression, Chi-square test*

*The initial approach will be to use a decision tree to determine which dealer level variables (size, region, segmentation...) are most significant related to a dealer’s likelihood to churn. We will also use other techniques to verify our findings.*

**Methodology:** *Key questions from ‘Analytics objective’ will be tackled in ascending order as outlined in ‘5.0 Timelines and deliverable section’.*

*We will start by identifying all dealers that were active in the first quarter of 2018. We will then define the response variable to be a 1 if they are still active, and 0 otherwise. We will build a decision tree based on this sample, and observe which variables are the most important in determining whether these dealers are still active. We can then repeat this analysis using a sample based on the dealers that were active in the second quarter of 2018. The idea is to check if the same variables are being identified as the most important drivers of churn, or if the importance of variables change as we get closer to the present day.*

**Output:** *The output will be a set of insights, rules and strategic recommendations that will help us to evaluate dealers based on likelihood to churn and positioning of sales-match.*

Methodology:

1. **Data Preprocessing:** Address any missing values, outliers, or inconsistencies in the dataset to ensure quality. Create new features or transform the existing ones to enhance predictive power of the model. (binning, normalizing)
2. **EDA:** Summarize the data to understand its structure, distribution, and key characteristics. Use various plots (histogram, boxplots, scatter plots) to show the relationship between variables and identify any patterns or anomalies.
3. **Model Building:** Build a logistic regression model to predict the likelihood of customer churn based on the identified predictor variables. Construct a decision tree to show the most influential variables that contribute to customer churn. This will help in understanding the decision rules that lead to churn. Random forest model can be used to improve the robustness and accuracy of prediction by aggregating the results of multiple decision trees.
4. **Model Evaluation:** Use k-fold cross-validation to assess the model’s performance. This involves splitting the data into k subsets and training/ testing the model k times. Evaluate the models using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC to determine their effectiveness in predicting churn.
5. **Insights:** In order to identify the most significant factors contributing to customer churn, I will interpret the model results and generate actionable insights that can inform the bank’s customer retention strategies

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** The audience will be the executives of the bank and the population are the customers in the dataset.

**Observation window:** The observation window is assumed to be in 2024.

**Inclusions:** Demographic Data: Age, gender, marital status, income level

Behavioral Data: Transaction history, product usage, account activity, customer service interactions, etc,

Account Information: Account type, balance, tenure with the bank, etc.

**Exclusions:** RowNumber, CustomerID, Surname

**Data Sources: (From Kaggle)** <https://www.kaggle.com/datasets/murilozangari/customer-churn-from-a-bank>

**Audience Level:** The analysis will be conducted at the individual customer level, with each customer representing a unique data point.

**Variable Selection:** Demographic Data: Age, gender, marital status, income level

Behavioral Data: Transaction history, product usage, account activity, customer service interactions, etc,

Account Information: Account type, balance, tenure with the bank, etc

**Derived Variables: To be Decided**

**Assumptions and data limitations:** **To be decided**

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| 1. *Churn rate being inflated by counting multiple contracts from the same customer as individual observations.* 2. Changes in customer behavior post-analysis 3. Data Integration Challenges 4. Regulatory and Compliance Issues | *Medium         Low          Medium                 Low* | *Moderate         Minimal          Moderate                 Minimal* | *If multiple contracts from the same customer are counted as individual observations, it could inflate the churn rate. This can lead to overestimation of the number of customers who are likely to churn.  Customer behaviour may change due to external factors (economic shifts, competitor actions) which may not be reflected through historical data used for the analysis.*  Integrating data from different sources (e.g., CRM systems, transaction records) can be challenging and time-consuming. Misalignment of data formats and inconsistencies can lead to integration delays and impact the project’s progress.     Any violation to regulatory and compliance requirements can result in legal repercussions and damage the bank’s reputation. |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Days** | **Date** |
| 1. | Define Business Problem | *Define objectives, clear business problem related to customer churn.* | *1* | *10/07/2024* |
| 2. | Prioritization | Prioritize key variables and define analysis approach | 1 | *11/07/2024* |
| 3. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Conduct EDA, identify duplicates, issues with spend data | 1 | *12/07/2024* |
| 4. | Meeting | Introduce the project centered on predicting customer churn at the bank. | 1 | *15/07/2024* |
| 5. | Project Plan Documentation | Document the plan which will become the skeletal of the project. | *2* | *15/07/2024* |
| 6. | Data pre-processing | In order to make the model accurate, data set can be modified by cleaning it in Python. | *1* | *17/07/2024* |
| 7. | Exploratory Data Analysis | EDA will be performed in Python to understand the data and its characteristics. | *1* | *18/07/2024* |
| 8. | Model Selection and Training | Build a logistic regression, construct a decision tree and create a random forest model and check their performance | *1* | *19/07/2024* |
| 9. | Story Board 1 | Refine storyboard based on feedback and prepare for final presentation | 3 | *22/07/2024* |
| 10. | Pilot | Implement pilot retention strategies based on insights. | *4* | *26/07/2024* |
| 11. | Delivery & sign-off | Final presentation to stakeholders | *1* | *16/08/2024* |