WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Name

Chris Winterowd

Student ID #

012286631

Date

10/23/24

1. **Project Overview**

This proposal focuses on building a machine learning model to help a bank predict which customers are likely to leave. The goal is to analyze customer data to find patterns that show who might stop using the bank’s services. With these insights, the bank can take action to keep at risk customers. This project aims to support the bank’s efforts to improve customer satisfaction and reduce losses by focusing on early intervention.

**A.1. Organizational Need**

The bank needs a program that can predict if a customer might leave so it can take action to help them stay. If the bank knows in advance which customers are at risk of leaving, it can offer them special deals, support, or personalized services. This will help improve customer satisfaction and reduce the chances of losing valuable customers. By keeping more customers, the bank can avoid losses and improve its reputation. The program will use customer data to find patterns that show which customers are most likely to leave, allowing the bank to act early and stay ahead of the competition.

**A.2. Context and Background**

The bank that hired me has been facing a growing issue with customer churn over the past few years. Many customers are leaving for their competition and they don’t know why.. This has made it harder for the bank to stay competitive and has caused a noticeable drop in profits. Customer service teams are overwhelmed, trying to understand why people are leaving, but they can only react after the fact, which is often too late to make a difference. The bank knows that keeping customers is much cheaper than finding new ones, so it’s looking for ways to act before customers decide to leave. My project will provide a solution by predicting which customers are at risk, giving the bank a chance to step in early with personalized offers or support to encourage them to stay.

**A.3. Outside Works Review**

**Outside Work #1:**

MODELING OF CHURN BEHAVIOR OF BANK CUSTOMERS USING LOGISTIC REGRESSION

[Research Link](http://ir.jkuat.ac.ke/bitstream/handle/123456789/1202/ndungu%20ann%20wachira%20-Mcs%20Applied%20Statistics-2013.pdf?sequence=1&isAllowed=y) (Will also be in sources cited)

**A.3 Summary of the Outside Work #1**

The study, "Modeling of Churn Behavior of Bank Customers Using Logistic Regression," examines the factors that influence customer churn in the banking industry. It identifies key drivers of churn such as bank reputation, service quality, price, and product offerings. The research demonstrates how logistic regression can be used to predict which customers are likely to switch banks, enabling managers to take action before they leave. The study highlights that understanding these factors helps banks improve customer retention and maintain profitability in a competitive environment.

**A.3.a How Outside Work #1 Relates to My Solution**

This study directly supports my project by showing the importance of early intervention in retaining customers. It emphasizes how banks can use customer behavior data to predict churn and take proactive steps to keep customers, such as improving service quality or offering incentives. My project aligns with this approach by building a model that identifies customers at risk of leaving, giving the bank an opportunity to act early and reduce losses. This proactive strategy, as highlighted in the study, ensures better customer retention and supports long-term profitability.

**Outside Work #2:**

Machine-Learning Techniques for Customer Retention: A Comparative Study

[Research Link](https://thesai.org/Downloads/Volume9No2/Paper_38-Machine_Learning_Techniques_for_Customer_Retention.pdf) (Will also be in sources cited)

**A.3 Summary of the Outside Work #2**

The paper, "Machine Learning Techniques for Customer Retention," explores various machine learning models to predict customer churn. It emphasizes the importance of using predictive analytics to identify at-risk customers and highlights that early identification allows businesses to take proactive steps to keep these customers. The study focuses on how customer behavior and transactional data are used to build models such as logistic regression, decision trees, and ensemble methods. The primary goal is to help organizations improve customer retention by acting before customers leave, which is more cost-effective than acquiring new ones.

**A.3.a How Outside Work #2 Relates to My Solution**

This paper highlights how using predictive models like logistic regression can transform customer retention strategies by providing businesses with better insight into customer behavior. Rather than reacting to churn after it happens, the model empowers the organization to predict churn and build targeted marketing campaigns or loyalty programs. My project aligns with this by enabling the bank to create personalized outreach efforts based on the predictions. This means the bank can tailor offers and services specifically to high risk customers, improving the chance of keeping them and building long term relationships.

**Outside Work #3:**

**Churn Prediction for Savings Bank Customers: A Machine**

**Learning Approach**

[Research Link](https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1314&context=jsap) (Will also be in sources cited)

**A.3 Summary of the Outside Work #3**

The paper, "Churn Prediction for Savings Bank Customers: A Machine Learning Approach," focuses on how machine learning models can be used to identify savings bank customers at risk of leaving. It examines how customer data, such as demographics, transactions, and account behavior, are utilized to predict churn. The study highlights that customer churn impacts profitability, emphasizing the need for strategies to retain customers. It also says that customer retention is critical for banks, particularly for high-value customers, because retaining existing customers is more profitable than trying to get new ones. It highlights that reducing churn can increase profits significantly and strengthen a bank’s competitive edge.

**A.3.a How Outside Work #2 Relates to My Solution**

This paper provides insights into how machine learning can prioritize high-risk customers for targeted retention efforts. My solution will use similar strategies to help the bank focus its resources and makes sure that the right customers are identified and offered personalized support. Instead of applying broad retention campaigns, the bank can use these predictions to reduce costs by concentrating efforts on those most likely to churn which should help them increase the chance a customer stays with their bank. This approach, highlighted in the paper, ensures that the bank’s retention efforts are both effective and cost efficient.

**A.4. Solution Summary**

The solution I propose is to build a machine learning model using logistic regression that predicts which customers are most likely to leave the bank. The model will analyze customer data such as behavior patterns, transaction activity, and demographics to classify customers into binary categories which are those likely to leave and those likely to stay. This binary classification allows the bank to focus its efforts on high risk customers by identifying them early.

With these predictions, the bank can proactively reach out to at risk customers with targeted offers, improved services, or personalized interventions to encourage them to stay. This solution aligns with the bank’s organizational need to reduce churn by acting early, instead of waiting until after customers have already left. By prioritizing customer retention, the bank can improve profitability and maintain stronger relationships with customers, reducing the impact of churn.

**A.5. Machine Learning Benefits**

The machine learning model will help the bank identify customers who are likely to leave early, giving the bank time to take proactive steps to retain them. This is important because keeping current customers is often much cheaper than trying to attract new ones. With accurate predictions, the bank can focus on at risk customers and offer them incentives, special services, or personalized support to encourage them to stay.

Another key benefit is that the model will help the bank use resources more efficiently. Instead of spending time and money on broad campaigns aimed at all customers, the bank can focus its efforts on the customers most likely to leave. This targeted approach ensures that the bank’s retention efforts are not only more effective but also more cost-efficient.

The model will contribute to the bank’s long term profitability and competitiveness. Reducing churn means fewer lost customers and more stable revenue, which strengthens the bank’s position in the market. Additionally, by offering better services to customers at risk of leaving, the bank can build stronger relationships, leading to increased customer satisfaction and loyalty.

**B. Machine Learning Project Design**

**B.1. Scope**

**In Scope**:

* Customer churn prediction using logistic regression to classify customers as likely to stay or leave the bank. The project will focus on simple, binary predictions to help the bank identify at-risk customers early.
* Data analysis using customer demographics, behavior patterns, and transaction history to improve prediction accuracy. This will ensure the model can spot patterns that indicate a higher chance of a customer leaving.
* Model training and testing with historical data to ensure the predictions are reliable and accurate. The data will be split into training and testing sets to evaluate how well the model works.
* Proactive retention strategies, such as identifying at-risk customers and recommending targeted offers or personalized support to retain them. The predictions will help the bank reach out to these customers with targeted solutions before they leave.

**Out of Scope**:

* Customer acquisition strategies or attracting new customers to the bank. This project will only focus on keeping existing customers from leaving.
* Integration with third-party platforms, such as marketing or CRM software, for automated customer outreach. The model will provide predictions, but further automation will not be part of this project.
* Advanced machine learning models like random forests or neural networks for churn prediction. This project will only use logistic regression for its simplicity and effectiveness.

**B.2. Goals, Objectives, and Deliverables**

**Goals**

* Build a machine learning model that helps the bank predict which customers might leave. This model will allow the bank to identify at-risk customers before they leave.
* Help the bank improve customer retention by acting before at risk customers leave. By taking action early, the bank can prevent losses and maintain long term relationships.
* Make the bank’s retention efforts more focused and efficient, leading to higher customer satisfaction and reduced losses. This will allow the bank to save both time and resources by targeting the right customers.

**Objectives**

* Train the logistic regression model using historical customer data to identify patterns of churn. This process will ensure the model learns to spot key behaviors linked to churn.
* Test the model’s accuracy to make sure the predictions are reliable and useful for decision-making. Testing will confirm that the model performs well with new or unseen data.
* Use the predictions to identify the top customers at risk of leaving and suggest actions the bank can take to retain them. This ensures the bank can act strategically to improve retention and profitability.

**Deliverables**

* A logistic regression model with at least 80% prediction accuracy in identifying at-risk customers. This ensures the model is reliable enough for the bank to confidently act on the predictions.
* A report detailing the model's development, including training methods, test results, and accuracy scores. This report will give the bank a clear understanding of how the model works and its performance.
* A list of the top 5% of at-risk customers identified by the model, with recommendations for retention strategies. This will help the bank focus its efforts on the customers most likely to leave, increasing the chance of keeping them.

**B.3. Standard Methodology**

Development will follow the SEMMA methodology.

**Sample**:

* A sample of the bank’s customer data will be selected, including demographics, transaction history, and account behavior. This ensures the model can be built efficiently without sacrificing the quality of the insights.

**Explore**:

* The data will be explored to identify trends and patterns that indicate which customers are likely to leave. This step provides a deeper understanding of which features are most important for making accurate predictions.

**Modify**:

* The data will be cleaned, formatted, and transformed to improve the model’s performance. This involves handling missing values, encoding categorical data, and scaling numerical data so the model can process it properly. Modifying the data ensures that the model can learn effectively and make accurate predictions from all relevant information.

**Model**:

* A logistic regression model will be developed using the prepared data. The model will be trained to detect patterns that indicate which customers are likely to churn. After training, the model will be fine tuned by adjusting parameters to improve performance, balancing accuracy and efficiency to ensure practical predictions for the bank’s use.

**Assess**:

* The model will be tested to evaluate how well it predicts churn. This step involves comparing predictions with actual outcomes from the test data to measure the model’s accuracy. If the predictions aren’t reliable, adjustments will be made to improve the model’s performance before deployment.

**B.4. Projected Timeline**

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| Task | **Start** | **End** | **Tasks** |
| 1: Data Collection and Preparation | February 5, 2025 | February 11, 2025 | Collect customer data, like demographics and transactions, and organize it for analysis. |
| 2: Data Exploration and Feature Engineering | February 12, 2025 | February 18, 2025 | Look for patterns in the data and pick key features, like account balance, that can improve predictions. |
| 3: Model Development | February 19, 2025 | February 25, 2025 | Build and train a logistic regression model using the prepared data to predict churn. |
| 4: Model Testing and Evaluation | February 27, 2025 | March 3, 2025 | Test the model’s accuracy and make improvements if needed to ensure it works well. |
| 5 | March 4, 2025 | March 7, 2025 | Submit the final model and report, with recommendations for how the bank can use the results. |

**B.5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Laptops/Computers | Used by team members for building, testing, and deploying the model. Each member needs a reliable machine. | $2,500 (x5) = $7,500 |
| Software (Anaconda) | Free software for local data preparation using Jupyter Notebooks. | $0 |
| Google Colab | Used for cloud-based model hosting, allowing faster computations and easy collaboration. | $0 |
| Power BI & Tableau | Visualization tools for reporting insights and presenting model results. | $100(Power BI) + $100(Tableau) = $200 |
| Work Hours (IT Team) | 5 IT members responsible for infrastructure setup, deployment, and maintenance, working 150 hours each. | 5 × (150 hrs × $70/hr) = $52,500 |
| Work Hours (Project Team) | 3 Project members working on data preparation, modeling, and testing, each working 160 hours | 3 × (160 hrs × $70/hr) = $33,600 |
|  | **Total** | $93,800 |

**B.6. Evaluation Criteria**

Describe the criteria used to evaluate and measure the success of the completed project.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Build a machine learning model to predict churn | The success will be measured by the model achieving at least 80% accuracy in identifying at-risk customers. |
| Improve customer retention by acting early | Success will be reflected in the bank retaining 5% of at-risk customers identified by the model. |
| Focus retention efforts on the right customers | Success will be measured by a reduction in unnecessary retention costs by at least 10% through targeted campaigns. |
| Ensure the model performs well with unseen data | The success will be evaluated based on the model’s ability to maintain accuracy with new or test datasets. |
| Provide actionable insights from predictions | Success will be shown by the bank’s ability to launch three targeted retention campaigns based on predictions. |
| Deliver detailed reports on the model’s performance | The success is achieved when a final report is submitted, including training methods, testing results, and accuracy scores. |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

The hypothesis of this project is that a logistic regression model can accurately predict whether a customer will stay with the bank or leave, using historical data. Since this is a binary classification problem, the model will categorize each customer as either "staying" or "leaving" based on patterns in their activity, product usage, and account history.

The hypothesis will be tested by splitting the dataset into training and testing sets. Performance metrics such as accuracy, precision and also the recall will be used to evaluate how well the model predicts these outcomes. If successful, the predictions will allow the bank to proactively target at risk customers, improving retention.

**C.2. Selected Algorithm**

The selected algorithm for this project is logistic regression, which is a type of supervised learning. This algorithm is well-suited for binary classification problems like predicting whether a customer will stay or leave the bank. Logistic regression works by analyzing the relationship between customer attributes (such as balance, age, or number of products) and the likelihood of churn, assigning probabilities between 0 and 1.

**C.2.a Algorithm Justification**

The logistic regression algorithm was chosen because it is simple, interpretable, and effective for binary outcomes. It allows the bank to easily understand which features are influencing customer churn, supporting actionable business decisions.

According to studies in the field of customer churn, logistic regression is commonly used due to it being able to produce accurate predictions (Ndungu, 2013). Also, logistic regression assigns a weight to each feature, showing how much it influences the prediction. This helps the bank see which factors, like balance or product usage, are most important in determining if a customer will leave.

**C.2.a.i. Algorithm Advantage**

An advantage of logistic regression is that it’s easy to set up and it’s also easy to understand. The model provides a probability score for each prediction, which helps the bank decide if a customer is more likely to stay or leave. By setting a simple threshold, the bank can quickly classify customers into "staying" or "leaving" categories. This makes it straightforward to interpret and act on the predictions without needing to use a bunch of complex calculations.

**C.2.a.ii. Algorithm Limitation**

One limitation of logistic regression is that it works best when the relationship between the data and the outcome is simple. If the customer behavior depends on many complicated factors working together, the model might not catch those patterns well. For example, if a combination of several subtle behaviors leads to churn, the model may miss it because it assumes a straight line relationship between the factors and the result. This could make it less accurate when dealing with complex data, meaning the predictions might not always be reliable.

**C.3. Tools and Environment**

For this project, I will use Windows 10 with Python as the primary language. Development will start locally using Jupyter Notebook through Anaconda and later be transferred to Google Colab for sharing and collaboration. The project will rely on scikit-learn for the logistic regression model, with pandas and matplotlib for data handling and visualization. Tools like Power BI and Tableau will support visual presentations of the results. No additional APIs or external services will be used, ensuring the solution remains straightforward and easy to maintain.

**C.4. Performance Measurement**

The performance of the machine learning solution will be measured using key metrics like accuracy, precision, and recall. Accuracy will show how often the model makes correct predictions, while precision and recall will help measure how well it identifies customers who are at risk of leaving. The model will also be tested with different sets of data to make sure it gives consistent and reliable results. Cross validation will be used to check that the model works well with new data and doesn’t just perform well on the training data.

**D. Description of Data Sets**

**D.1. Data Source**

The data used for this project will be extracted from the bank’s internal customer management system. It will include historical information on customer transactions, account activity, tenure, demographic data, and interactions with the bank’s services. This internal dataset will provide a reliable foundation for building the predictive model since it reflects actual customer behavior and patterns when the customer is interacting with the bank. Using this data ensures that the model is made to the bank’s specific needs and accurately identifies customers at risk of leaving.

**D.2. Data Collection Method**

The data will be collected through a combination of automated and manual processes. Automated collection will involve APIs connected to the bank’s internal systems, pulling in data such as transactions, account activity, and customer profiles on a daily basis. Additionally, staff members will manually input data related to specific interactions, like in person meetings or customer service calls, to capture behavioral insights that might not be reflected through records from a customer using a program. Scheduled data backups will also ensure that no critical information is lost during the collection process.

**D.2.a.i. Data Collection Method Advantage**

One key advantage of this approach is the efficiency it brings by automating most of the data collection. APIs ensure a continuous flow of real time information, which minimizes the need for a lot of manual updates. This allows the machine learning model to quickly access the latest customer data, helping the bank stay proactive in identifying customers at risk of leaving.

**D.2.a.ii. Data Collection Method Limitation**

A limitation of this method is the potential reliance on consistent system integration. If the APIs encounter technical issues or if staff members miss inputting data manually, there could be gaps in the information collected. These gaps might reduce the model’s effectiveness, as incomplete data can lead to less accurate predictions. Regular monitoring and system maintenance will be essential to prevent disruptions in the data collection process.

**D.3. Quality and Completeness of Data**

To prepare the data for the logistic regression model, any missing values will be handled by either filling them with averages or removing incomplete records. Outliers that could disrupt predictions will be identified and managed to ensure the data remains consistent. The dataset will also be checked for errors, such as duplicates or incorrect entries, which will be corrected or removed. Additionally, the data will be converted into a format that works well with the algorithm, ensuring each feature is numeric or transformed appropriately so that it can be used for binary classification. This ensures that the model can accurately identify patterns and produce reliable predictions.

**D.4. Precautions for Sensitive Data**

It’s important to handle sensitive data carefully to protect privacy and avoid misuse. The behaviors the team will use starts by only allowing the specific team members who need access to view the data, also making sure the data is encrypted when stored and shared. Another important thing is to use secure communication methods like encrypted emails. It’s also essential to remove or hide personal information to keep the data anonymous whenever possible. Regularly training the team on data protection practices would also be good, this would help everyone follow the same guidelines. Also, following general data protection regulations helps maintain trust and ensures legal compliance when managing customer information.

**References**

"Churn Prediction for Savings Bank Customers: A Machine Learning Approach." Journal of Statistics Applications & Probability, https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1314&context=jsap

Sabbeh, Sahar F. "Machine-Learning Techniques for Customer Retention: A Comparative Study." International Journal of Advanced Computer Science and Applications, vol. 9, no. 2, pp. 273-276, https://thesai.org/Downloads/Volume9No2/Paper\_38-Machine\_Learning\_Techniques\_for\_Customer\_Retention.pdf

Ndung'u, Anne Wanjira. Modeling of Churn Behavior of Bank Customers Using Logistic Regression. Submitted in Partial Fulfillment of the Requirements for

Award of a Masters of Science in Applied Statistics, Jomo Kenyatta University of Agriculture and Technology, http://ir.jkuat.ac.ke/bitstream/handle/123456789/1202/ndungu%20ann%20wachira%20-Mcs%20Applied%20Statistics-2013.pdf?sequence=1&isAllowed=y