**Enhancing Churn Prediction for the Banking Sector: Machine Learning and Feature Engineering**

**Interm Progress Report**

**STUDENT ID: 21083247**

**STUDENT NAME: MD Mahfuj Khan**

**COURSE: MSc Data Science and Analytics**

**Supervised By: Daniel Onah**

Table of Contents

[**Background Study:** 3](#_Toc165865170)

[**Aim and Objectives** 3](#_Toc165865171)

[**Research Questions:** 4](#_Toc165865172)

[**Dataset Information:** 4](#_Toc165865173)

[**Literature Review:** 5](#_Toc165865174)

[**Ethical Considerations:** 6](#_Toc165865175)

[**Methodologies**: 6](#_Toc165865176)

[**Project Progress:** 7](#_Toc165865177)

[**Project Plan:** 7](#_Toc165865178)

[**References:** 9](#_Toc165865179)

[**Appendices** 10](#_Toc165865180)

# **Background Study:**

In the driven landscape of the banking sector, customer churn, which refers to customers leaving a service, establishes a significant challenge. Churn impacts a bank's revenue flows and hold back its ability to grow sustainably. Traditional statistical methods have proven insufficient in effectively predicting churn, as they often cannot handle the complexity and non-linear patterns found in customer behavior data.

This is where Machine Learning (ML) steps in, offering a more dynamic and insightful approach. ML algorithms can process vast amounts of data and recognize complicated patterns that might indicate a likelihood of churn. However, these models come with their own set of challenges, particularly in dealing with class imbalance—a scenario where the number of churned customers is substantially less than the non-churned, which is common in churn datasets. This imbalance can skew the model's performance, leading to an over-prediction of the majority class and under-prediction of the minority class.

To address these issues, the project investigates advanced ML techniques, such as combination methods, that have the potential for better performance by combining the strengths of various learning algorithms. The focus is also on evaluating strategies to mitigate class imbalance, like resampling methods that either undersample the majority class or oversample the minority class to achieve a more balanced dataset. These strategies include the use of Synthetic Minority Over-sampling Technique (SMOTE), which generates artificial examples rather than simply duplicating minority class instances.

The project places highlighting on understanding the drivers of churn. Through feature engineering, Project aim to create new predictive variables from existing data that can provide deeper insights. For instance, customer demographic factors, transaction history, product usage, and engagement levels are transformed into metrics that the ML models can utilize to recognize patterns leading to churn.

Given the variety of potential predictors, diverse ML models including logistic regression, decision trees, and random forests are evaluated. Hyperparameter tuning and model comparison are conducted to pinpoint the most effective model, with a balanced consideration of precision and recall—metrics that are critical in the context of imbalanced classes.

By delving into predictive accuracy and model interpretability, the project's ambition is to produce a churn prediction model that not only accurately identifies at-risk customers but also informs targeted retaining strategies. The success of this endeavor can equip banking institutions with a proactive tool, transforming the way they manage customer relationships and mitigate churn.

# **Aim and Objectives**

**Aim**: To Build a robust customer churn prediction model for banks, prioritizing balanced performance with high precision and recall, while identifying key churn drivers and addressing data imbalance.

**Objectives**:

Prepare: Clean and analyze customer churn data.

Model: Develop and compare machine learning models for imbalanced classification, optimizing hyperparameters.

Balance: Address class imbalance through resampling, cost-sensitive learning, and other techniques.

Features: Identify key churn drivers using feature importance and create new informative features.

Evaluate & Analyze: Assess models with robust metrics and derive actionable insights for the banking industry.

# **Research Questions:**

1. Which machine learning algorithm, among XGBoost, Random Forest, and Logistic Regression, exhibits the optimal balance between precision and recall for predicting customer churn within the context of the given dataset?
2. What are the most influential features driving customer churn decisions, and are there visible variations in feature importance across the evaluated machine learning models?
3. Does customer age, tenure, and the combination of bank products utilized have a statistically significant impact on the likelihood of churn?
4. Do significant differences in churn rates exist across geographic regions, and can these differences be attributed to variations in demographic or account-related features?
5. Does the inclusion of the Credit Score feature improve churn prediction performance, and how does its predictive power compare to traditional indicators such as account balance or tenure?

# **Dataset Information:**

Dataset Overview:

The dataset was taken from an open source platfrom Kaggle. The dataset contains customer data from a bank, with each row representing a customer profile and various attributes collected during their tenure with the bank.

URL: <https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling>

Common Features:

CustomerId: A unique identifier for the customer.

Surname: The last name of the customer.

CreditScore: score assigned to the customer based on their credit history.

Geography: Regional sector to which the customer belongs.

Gender: Gender of the customer.

Age: Age of the customer.

Tenure: How many years the customer has been with the bank.

Balance: The current balance in the customer's account.

NumOfProducts: Number of products the customer has with the bank.

HasCrCard: Indicates whether the customer has a credit card with the bank or not.

IsActiveMember: Indicates whether the customer is an active member or not.

EstimatedSalary: Estimated salary of the customer.

Exited: Target variable indicating whether the customer has churned (left the bank) or not.

Preprocessing Steps:

Removing identifiers and irrelevant information like CustomerId and Surname.

Converting categorical variables into numeric using techniques like one-hot encoding.

Scale continuous variables to ensuring that they contribute equally to model training.

Handling missing values if present, either by imputation or by removing the affected records.

Purpose for Analysis:

The primary goal with this dataset is to predict the Exited feature, which indicates whether a customer has churned.

Understanding the factors that contribute to customer churn can help the bank to implement targeted holding strategies.

# **Literature Review:**

In the rapidly evolving banking sector, customer retention is crucial for sustained growth and profitability. Verbeke et al. (2012) delve into churn prediction within the telecommunications sector, initiating a profit-driven data mining approach that has inspired similar strategies in banking. They highlight the economic impact of churn and the value of predictive analytics in modifying its effects. Larivière and Van den Poel (2005) further this concept by demonstrating the application of random forests and regression forests in predicting customer retention and profitability, highlighting the superiority of these methods over traditional statistical techniques.

Exploring the efficiency of classifier ensembles in real-world scenarios, Oza and Tumer (2008) reveal that such approaches, which integrate multiple models, can significantly outperform single-model systems. These ensembles, by leveraging the diversity of individual classifiers, improve prediction accuracy—principles highly applicable to churn prediction models in banking. The challenge of imbalanced datasets, common in churn prediction, where the number of customers who churn is typically much lower than those who do not, is addressed by He and Garcia (2009). They propose methods for learning from imbalanced data, which is instrumental in developing more effective churn prediction models.

The determining work of Freund and Schapire (1997) introduces a boosting algorithm that iteratively corrects the mistakes of weak classifiers, forming a strong predictive model. Boosting algorithms have since become a foundation in churn prediction for their ability to improve the accuracy of models significantly. To evaluate the performance of such models, Fawcett (2006) suggests ROC analysis, which has been widely adopted as a measure of model effectiveness, particularly in differentiating between churned and retained customers.

In addressing class imbalance specifically, García et al. (2012) focus on evolutionary-based selection of generalized instances, akin to the Synthetic Minority Over-sampling Technique (SMOTE). This method synthesizes new samples from the minority class, ensuring that the model does not become biased towards the majority class. This technique is particularly relevant given the essential class imbalance in churn prediction datasets. Kohavi et al. (1995) contribute to this discussion with their examination of cross-validation and bootstrap methods, which are critical for accuracy estimation and model selection in predictive churn models.

Vafeiadis et al. (2015) present a comparative study of machine learning techniques for churn prediction, underscoring the importance of choosing the right algorithm for specific dataset characteristics. Lastly, Amin et al. (2019) approach churn prediction in the telecommunications industry from a big data perspective, utilizing machine learning to manage and analyze vast datasets—a methodology equally applicable to the banking sector.

These studies collectively underscore the necessity of advanced analytical techniques in addressing churn. They advocate for ensemble methods, sophisticated data balancing techniques, and rigorous model evaluation metrics as pivotal components in the development of churn prediction models. As such, they form the theoretical and methodological strength of current efforts to enhance churn prediction accuracy in the banking sector.

# **Ethical Considerations:**

Ethical Considerations:

Privacy: Ensuring customer data is used responsibly, without violating on individual privacy rights. Data should be anonymized to protect identities.

Bias and Fairness: The model must be free from biases that could lead to unfair treatment of certain customer groups based on gender, age, or ethnicity.

Transparency: There should be clarity in how the model makes predictions, and customers should be informed about how their data is being used.

# **Methodologies**:

Data Acquisition (Week 1):

The dataset was sourced from Kaggle, specifically chosen for its relevance to churn prediction in the banking sector. This dataset includes customer profiles and their interaction data with the bank. The dataset contains all necessary features like customer demographics, account details, and churn status.

Data Cleaning (Week 2):

Initial data was inspection involved identifying missing values, irrelevant features, and data types to prepare for preprocessing. Features that do not influence churn prediction, such as CustomerId, Surname, and RowNumber, were removed to simplify the dataset.

Handling Missing Data: Any missing values identified during the inspection were addressed either by imputation or removal, although specific details depend on the dataset's characteristics as observed.

Data Preprocessing (Week 3):

Feature Scaling: Numerical features like Age, Tenure, Balance, and others were standardized using scaling techniques to ensure they contribute equally to model performance.

Encoding Categorical Variables: Categorical features such as Geography and Gender were transformed using one-hot encoding to convert them into a machine-readable format.

Feature Engineering (Week 4 - 5):

New Feature Creation: Based on insights from the exploratory data analysis, new features were engineered to enhance the model's predictive power. This includes interaction terms, aggregated features, or derived attributes like customer engagement scores.

# **Project Progress:**

Week 1: Sourced a churn dataset from Kaggle, featuring customer profiles and banking behaviors.

Week 2: Cleaned the data by removing identifiers and irrelevant variables, ensuring quality for analysis.

Week 3: Preprocessed data through scaling numerical features and encoding categorical variables, preparing it for machine learning.

Week 4 - 5: Conducted feature engineering to enhance model inputs by creating meaningful variables from existing data.

# **Project Plan:**

Week 5-6: Plan to evaluate model performance using cross-validation, fine-tune hyperparameters, and analyze feature importance.

Week 7-8: Intend to interpret model results, delve into model explainability (possibly using SHAP values), and draft initial findings.

Week 8-9: Aim to validate findings with statistical tests, investigate potential bias.

Week 10-15: Preparing the Documentatio

# **References:**

Verbeke, W., Martens, D., Mues, C., & Baesens, B., 2012. 'New insights into churn prediction in the telecommunication sector: A profit driven data mining approach.' European Journal of Operational Research, 218(1), pp.211-229.

Larivière, B. & Van den Poel, D., 2005. 'Predicting customer retention and profitability by using random forests and regression forests techniques.' Expert Systems with Applications, 29(2), pp.472-484.

Oza, N.C. & Tumer, K., 2008. 'Classifier ensembles: Select real-world applications.' Information Fusion, 9(1), pp.4-20.

He, H. & Garcia, E.A., 2009. 'Learning from imbalanced data.' IEEE Transactions on Knowledge and Data Engineering, 21(9), pp.1263-1284.

Freund, Y. & Schapire, R.E., 1997. 'A decision-theoretic generalization of on-line learning and an application to boosting.' Journal of Computer and System Sciences, 55(1), pp.119-139.

Fawcett, T., 2006. 'An introduction to ROC analysis.' Pattern Recognition Letters, 27(8), pp.861-874.

García, S., Derrac, J., Cano, J., & Herrera, F., 2012. 'Evolutionary-based selection of generalized instances for imbalanced classification.' Knowledge-Based Systems, 25(1), pp.3-12.

Kohavi, R. et al., 1995. 'A study of cross-validation and bootstrap for accuracy estimation and model selection.' In: IJCAI International Joint Conference on Artificial Intelligence, pp.1137-1145.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., & Chatzisavvas, K.C., 2015. 'A comparison of machine learning techniques for customer churn prediction.' Simulation Modelling Practice and Theory, 55, pp.1-9.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., & Hussain, A., 2019. 'Customer churn prediction in telecom using machine learning in big data platform.' Journal of Big Data, 6(1), p.28.

# **Appendices**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

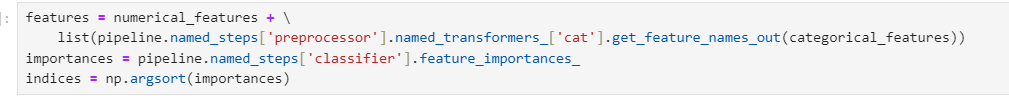
Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated



A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated