Customer Attrition Forecasting

Ashutosh Kalsi
ashutosh.kalsi@colorado.edu
CSCI 6502-001

Meghana Vasanth Shettigar
Meghana.VasanthShettigar@colorado.edu
CSCI 6502-001

Nitish V.S. Ramakrishnan
NitishVenkatesh.SeptankulamRamakrishnan@colorado.edu
CSCI 6502-001

Rajagopal Anandan
Rajagopal.Anandan@colorado.edu
CSCI 6502-001

ACM Reference Format:
Ashutosh Kalsi, Meghana Vasanth Shettigar, Nitish V.S. Ramakrishnan, and Rajagopal Anandan. 2023. Customer Attrition Forecasting. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Customer Churning is a significant issue in all industries. It is imperative to find client information and infer the knowledge that causes churn. The degree of customer inactivity or disengagement observed over a given time is defined as churn. Like with all industries, the banking sector is also affected by this issue. The banking industry has evolved significantly over the years, but customer churn remains a critical issue that many banks with large customer bases have yet to address despite having access to vast amounts of data. Retaining existing customers and increasing their lifetime value is essential even for the stock market prediction, yet banks often struggle to identify customer churn in advance, particularly in the absence of clear customer feedback. For instance, in the case of High Net-Worth customers, it is useful to define churn based on the rate of decline of assets over a specified period. There could be an instance where a customer may be highly active in terms of account operations but has effectively pulled out more than 50 percent of her assets in the last six months. Our project focuses on the issue of customer churn in the banking industry and the need for early and accurate churn prediction to retain customers and increase their lifetime value. We are planning to use machine learning and data science techniques to extract early warning signs of potential customer disengagement from the existing data. This approach can help gain a better understanding of the customer's perspective and potentially cross-sell complementary products to strengthen the relationship.

2 RELATED WORK

This section presents a summary of past work on churn prediction. In [1] the authors have proposed to predict customer churn in the banking sector with the help of machine learning techniques. They have used Support Vector Machine(SVM), K-Nearest Neighbours(KNN), Decision Tree, and Random Forest algorithms to create multiple models and evaluate their performance primarily on the accuracy of the predictions. To improve the model's performance, a feature selection process was applied to find the more relevant features from the dataset. Two different feature selection processes were used namely, Minimum Redundancy and Maximum Relevance(mRMR) and Relief. Using the feature selection process the number of features was reduced from 14 in the original dataset to 6 in the processed data. The original dataset has about an 80:20 ratio between non-churn customers and churn customers. To deal with this imbalance in class ratios, the authors have performed Oversampling technique to create a balanced dataset. They have used random oversampling by resampling the minority class (customers who churned). The accuracy of the resulting models using the mRMR feature selection process is listed in Table 1. It can be seen that Random forest performs the best after oversampling with an accuracy of 92.95%. The SVM model performs the worst with an accuracy of 69.96%.

Table 1: Accuracy Table using mRMR feature

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	83.97	82.57
SVM	79.63	69.96
DT 3	78.32	91.73
RF 4	83.66	92.95

The accuracy of the resulting models using the Relief feature selection process is listed in Table 2. It is observed that Random forest performs the best after oversampling with an accuracy of 92.19%. The SVM model performs the worst with an accuracy of 69.53%.

Table 2: Accuracy Table using Relief feature

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	82.15	80.99
SVM	79.63	69.53
DT 3	77.61	90.74
RF 4	81.75	92.19

Based on the results displayed above, the authors have concluded the Random Forest model to be the best-performing model after oversampling.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The authors of [2] have proposed to predict customer churn in Banking, Insurance, and Telecommunication sectors with the help of machine learning techniques. The proposed framework is composed of models using Linear Discriminant Analysis (LDA), Logistic Regression (LR), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Support Vector Machine(SVM), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP). The results for the banking models are described below. The authors have evaluated the models using Accuracy (AC), Precision (PR), f-Measure (FM), and Recall (RC). Accuracy is used as the primary measure of evaluation. The authors have used a feature reduction process to improve the prediction accuracy of the system. Principal Component Analysis (PCA) is applied for feature reduction and the number of features has been reduced from 21 in the original dataset to 15 in the processed data. The accuracy of the resulting models without the feature reduction process is listed in Table 3. It can be seen that Random forest performs the best after oversampling with an accuracy of 86.38%. The Decision Tree model performs the worst with an accuracy of 77.68%.

Table 3: The results of all classification techniques without feature reduction

Banking	Evaluation Metrics			
Models	AC	FM	PR	RC
LDA	81.83	83.14	81.25	81.60
LR	83.51	85.10	76.23	85.02
k-NN	84.23	87.35	81.37	88.55
NB	80.79	88.23	83.93	84.68
SVM	82.46	82.84	85.00	87.02
DT	77.68	83.10	80.59	85.94
RF	86.38	87.45	87.65	86.06
MLP	84.60	83.48	85.23	85.90

The accuracy of the resulting models using the PCA feature reduction process is listed in Table 4. It can be seen that Random forest performs the best after oversampling with an accuracy of 89.93%. The Decision Tree model performs the worst with an accuracy of 80.38%.

Table 4: The results of all classification techniques after PCA feature reduction

Banking	Evaluation Metrics			
Models	AC	FM	PR	RC
LDA	84.42	85.70	84.36	89.45
LR	85.57	87.43	86.90	90.13
k-NN	86.30	85.30	84.21	90.51
NB	81.22	83.01	82.57	87.24
SVM	85.55	86.40	85.68	88.00
DT	80.38	80.27	79.26	85.02
RF	89.93	90.51	89.75	93.57
MLP	87.40	88.13	87.59	90.45

Based on the results displayed above, the authors have concluded the Random Forest model to be the best-performing model.

3 PROPOSED WORK

Our proposed work summarizes our intended plan of action for our project. The work can be broken down into two broad tasks: Model implementation and Model Deployment. In addition, information on the dataset is added before the implementation for better understanding.

Dataset: The dataset we currently plan on using was obtained from Kaggle. This dataset has redacted customers' personal information such as Name, address, card information, etc. With relation to the bank and the clientele, we have information on the customers' past transactions with the bank, unique ID, credit score, estimated salary, account balance, and whether the customer has exited the bank or not. Miscellaneous information includes gender, last name (which will be redacted for privacy), age, and geographical location.

Model Implementation: We find it crucial to understand and prepare the dataset. The first task is to work on explanatory data analysis. Through visualizations, we aim to have a thorough insight into which factor(s) are responsible for the churn of customers. We use this to establish relations/associations between data features and customers' proclivity to churn and build a classification model to predict whether the customer will leave the bank or not. We can derive from the given problem that this is a classification problem. However, before moving forward with implementation, it is imperative to clean and feature engineer our dataset to set it up for proper implementation. Based on prior/ related work, our intention is to use algorithms that haven't been previously used. This broadly consists of neural network algorithms and we plan to use a range of algorithms from Convolutional Neural Networks to Long Short-Term Memory. Our intention is to squeeze maximum performance with this operation. In addition, we aim to optimize the machine-learning model by tuning the hyperparameters. Figure 1 is a pictorial representation (flowchart) of an approach to generate a standard machine learning model. We plan to utilize this methodology to generate the optimal solution to the problem at hand.

Model Deployment: Our current plan of action is to deploy our model using Amazon Web Services (AWS). We are planning to convert ML application to Flask application to do AWS deployment. We will deploy the application on Amazon Web Services (AWS) using Terraform, EC2 load balancer, S3, ECS cluster, ECR, and Docker image. AWS is a cloud-based platform that offers a wide range of services for computing, storage, and networking. It will help in easy deployment and scaling of the application. We will use Terraform as an infrastructure as code to manage and provision resources on AWS. It reduces the complexity and makes the deployment process faster and more efficient. AWS EC2 Load Balancer is a service which will be helpful in distributing the incoming traffic across multiple EC2 instances to ensure high availability, fault tolerance, and scalability. S3 is a simple storage service that provides scalable object storage for data backup and recovery. ECS cluster is an Amazon Elastic Container Service that manages Docker containers and their infrastructure. Using ECR (Elastic Container Registry), a fully-managed Docker container registry, it will be easy for us to store, manage, and deploy Docker container images. Figure 2 shows our present deployment plan for the project.

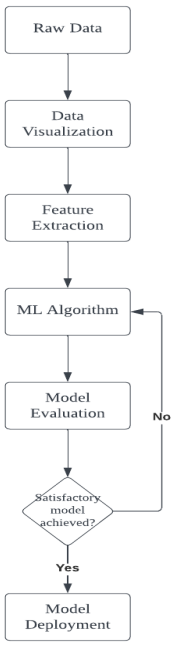

Figure 1: Planned Approach for Machine Learning Model

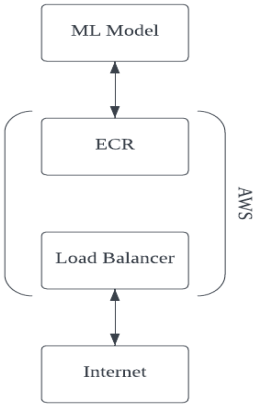

Figure 2: Planned approach for deployment

4 HOW TO EVALUATE

Evaluation metrics evaluate the model's performance. In our case, our plan is to use two checkpoints: one at model evaluation, and one at the end-stage of model deployment. For the model evaluation, we are currently planning to use the same metrics that were used in prior work i.e. Accuracy, F1 score, Precision, and Recall score. Our plan of action for deployment is to verify/ audit the model quality, correctness, and verify the integrity of the setup.

5 MILESTONES

Figure 1 and 2 are flowcharts depicting the methodology/ plan of approach. Currently, our milestones align (similarly) with the aforementioned flowcharts. Following is the list of the Milestones.

Problem Statement and Project Scope - Clearly define the business problem to be solved and the scope of the project.
Data Collection and Preparation - Identify and collect relevant data, clean and preprocess the data, and transform it into a format suitable for analysis.
Exploratory Data Analysis (EDA) - Conduct EDA to understand the underlying patterns, correlations, and trends in the data.
Feature Engineering - Create new features or extract relevant features from the data that can be used to build predictive models.
Model Selection and Development - Choose appropriate models and develop them using the prepared data. Evaluate model performance and refine models based on feedback.
Model Deployment and Monitoring - Deploy the final model and monitor its performance over time to ensure it remains accurate and effective.

6 REFERENCES

[1] M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction In Banking," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1196-1201
[2] Kilimci, Zeynep. (2022). The Effectiveness of Homogeneous Classifier Ensembles on Customer Churn Prediction in Banking, Insurance, and Telecommunication Sectors. International Journal of Computational and Experimental Science and Engineering.