Presentation Recording:

<https://drive.google.com/file/d/1hXvSirwSQyeQSro2eVJ42Tq85ve_0V75/view?usp=sharing>

**DAT565E Final Report**

**Olin Business School**

**Group 14：**[**Cyrus Cai**](https://wustl.instructure.com/groups/116463/users/93816) **- 517602，** [**Angel Liang**](https://wustl.instructure.com/groups/116463/users/95950) **- 519561，** [**Yongfan Zhao**](https://wustl.instructure.com/groups/116463/users/95945) **(MSQF)**

**Abstract**

Customers who stop using a company’s services are called "customer churn," they are a big problem for businesses in the financial sector, especially banks. The goal of this project is to use a bank dataset, including data like credit scores, to build a model that can predict whether a customer will leave. With the model's prediction, banks can set up specific plans to keep high-risk customers.

The dataset has 10,000 records. Before training our model, the data will be cleaned up. Different oversampling and undersampling methods will be used to fix the class mismatch. These methods can get helpful information that can help banks improve their strategies to keep customers by predicting and lowering the risks of customer churn. The best model shows a validation accuracy of 81.44%, showing that it is possible for banks to accurately predict whether a customer is going to leave.

**Introduction**

Customer churn is a significant concern for businesses in the financial sector, particularly banks. It represents the loss of customers who choose to close their accounts or discontinue using the bank's services. The ability to predict and mitigate churn can directly impact a bank's profitability by allowing it to implement effective customer retention strategies. In this project, we aim to develop a predictive model for customer churn using a dataset from a bank, which includes customer demographic, financial, and activity data.

We analyze this dataset to uncover patterns and factors most strongly correlated with churn, such as customer credit scores, balances, and engagement with the bank’s services. By applying deep learning techniques, we aim to create a model that can accurately predict the likelihood of a customer leaving. With these insights, banks can proactively retain high-risk customers through personalized offers and services, thus enhancing customer satisfaction and reducing churn.

The dataset for this project contains 10,000 records, with variables capturing customer demographics, financial data, and usage behaviors. We will preprocess and normalize the data before training a model using a Multilayer Perceptron (MLP) architecture to predict churn. Our approach also includes exploring the key variables that drive churn predictions, providing actionable insights for the bank.

**Literature Review**

Customer churn prediction has become a critical focus for industries, especially in sectors like banking and finance, where retaining customers is often more cost-effective than acquiring new ones. The development of machine learning and deep learning techniques has made it possible to predict customer churn with increasing accuracy. This literature review explores the most relevant studies that have shaped the design and implementation of churn prediction models, focusing on class imbalance solutions, model architectures, and application challenges in the financial sector.

One of the key challenges in churn prediction is managing imbalanced datasets, where the number of customers who churn is significantly smaller than those who do not. Mujahid et al. (2024) investigated various oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), to address the issue of class imbalance. Their work demonstrated that oversampling can significantly improve model accuracy by generating synthetic data points for the minority class, thereby preventing models from being biased toward the majority class​(s40537-023-00721-8). This approach directly influenced our decision to use SMOTE in our project, where we aimed to improve the prediction of churned customers.

Deep learning has emerged as a powerful tool for classification tasks, including churn prediction. Huang et al. (2020) provided an extensive review of deep learning applications in finance, highlighting the utility of neural networks in modeling complex relationships within financial data​(s11782-020-00082-6). Neural networks, particularly fully connected neural networks (FCNN), are known for their ability to handle structured data and uncover nonlinear patterns. For our project, we implemented an FCNN model with multiple layers of ReLU activation and a sigmoid output layer, following similar practices outlined in the literature. The use of techniques like dropout and L2 regularization, as explored by Huang et al. (2020), was critical in mitigating overfitting issues that we encountered during model training.

Convolutional neural networks (CNNs) have been widely used for tasks like image recognition, where local patterns are crucial. Alzubaidi et al. (2021) reviewed the evolution of CNN architectures and their applications beyond traditional image data, including structured financial datasets​(s40537-021-00444-8). However, in our project, we found that CNNs did not perform better than FCNNs, which aligns with findings in domains where spatial hierarchies are not as relevant. This highlights the importance of selecting the appropriate model architecture based on the nature of the data.

Furthermore, Suh (2023) applied machine learning models to predict customer churn in the home appliance rental business, focusing on real-world data rather than benchmark datasets​(s40537-023-00721-8). This practical approach to customer churn prediction closely mirrors our goal of developing a predictive model using real-world customer data from a bank. The emphasis on evaluating models through performance metrics like accuracy, F1 score, and AUC (area under the curve) was crucial in our assessment of model performance, where we achieved an AUC score of 0.93 using our best-performing FCNN.

Najafabadi et al. (2015) addressed the broader challenges of applying deep learning techniques to big data analytics, particularly in terms of computational efficiency and model scalability​(s40537-023-00721-8). Their findings on regularization techniques, such as dropout, L2 regularization, and early stopping, informed the strategies we employed to optimize our model. While our dataset was smaller in comparison to big data applications, the principles of managing overfitting and improving model generalization were directly applicable.

In summary, the studies reviewed here informed key decisions in our project, from managing class imbalance with SMOTE to selecting and optimizing model architectures. Our approach was strongly influenced by existing research on both machine learning and deep learning, which guided the successful application of these techniques in predicting customer churn for a bank. Future work will build upon these foundations by exploring alternative architectures like recurrent neural networks (RNNs) and further improving model performance through advanced feature engineering and real-time data integration.

**Problem Description**

Customer churn poses a substantial challenge for the banking industry, as losing customers can significantly impact both short-term revenues and long-term relationships. To address this issue, banks must not only understand the reasons behind churn but also develop predictive models that can forecast which customers are at high risk of leaving.

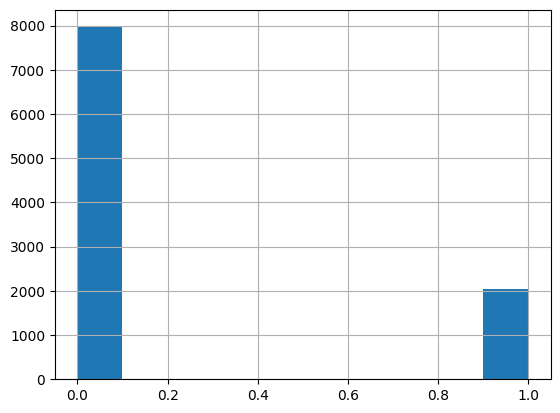
The goal of this project is to build a model that can accurately predict customer churn based on various factors such as demographics, financial information, and engagement with the bank’s services. By leveraging data-driven insights, the bank can intervene proactively, offering personalized retention strategies to at-risk customers. Specifically, this project will identify the key drivers of churn and use machine learning techniques to predict which customers are likely to close their accounts. The predictions will enable the bank to take targeted actions aimed at retaining high-risk customers and improving overall retention rates.

**Database Background and Data Preprocessing**

The dataset used for this project consists of 10,000 customer records obtained from a bank. It contains the following features:

* Customer demographics: Geography, Gender, Age.
* Financial information: CreditScore, Balance, EstimatedSalary.
* Customer activity: Number of products used, whether the customer holds a credit card (HasCrCard), and whether they are actively using the bank’s services (IsActiveMember).
* Tenure: The number of years the customer has been with the bank.
* Churn indicator: Exited (1 if the customer churned, 0 if not).

Before creating the model, we must convert all data formats to integers. We can check the format of each column using df.info, where the columns Surname, Geography, and Gender are of the object type. We need to process these columns accordingly. The Surname column doesn't contribute to our model analysis, so we will drop this column. For the Geography column, which includes values such as France, Germany, and Spain, we will use dummy variables to distinguish them. For the Gender column, we will use 1 and 0 to represent male and female, respectively. We also used df.Exited.hist() to understand the distribution of results within the 10,000 data points. However, we found that the distribution of these 10,000 data points is not very balanced, as shown in the figure below.

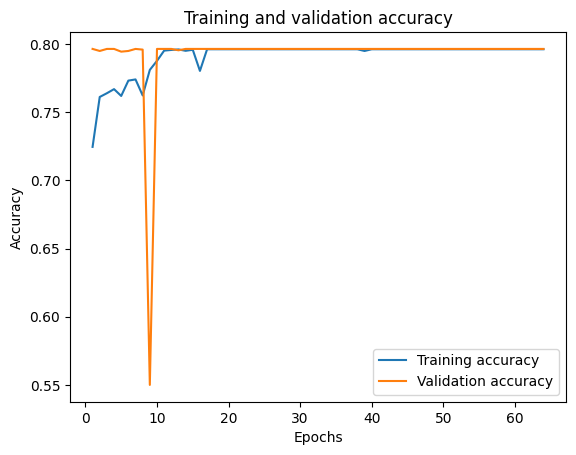


Since the model encounters far more majority class (class 0) samples than minority class (class 1) samples during training, it may learn to favor predicting the majority class. This could lead to a model with seemingly high accuracy but poor predictive performance for the minority class.

The data will be preprocessed before model training, including normalizing numerical variables, encoding categorical data, and handling missing values. The processed data will then be used to train a deep learning model (Multilayer Perceptron, MLP) to predict whether a customer is likely to churn. The results of this model will offer actionable insights into the factors that contribute most to churn, helping the bank to implement better retention strategies.

**Model**

We initially used a simple 4-layer model with 64, 32, 16, and 1 units, where the activation functions for the first three layers were ReLU, and the last layer used a sigmoid activation function. However, from the train accuracy and test accuracy graphs, we noticed significant fluctuations in the training results. Therefore, in subsequent training, we added a layer to the model.

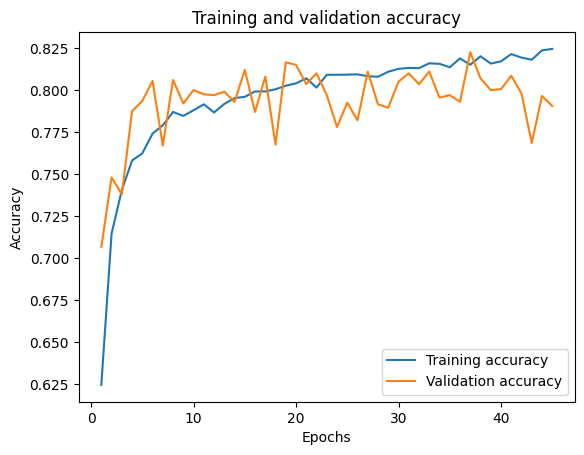
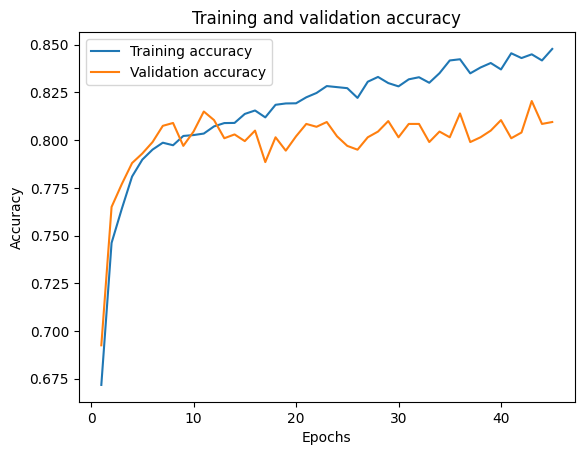


However, our test accuracy remained constant during this training, as shown in the figure above.

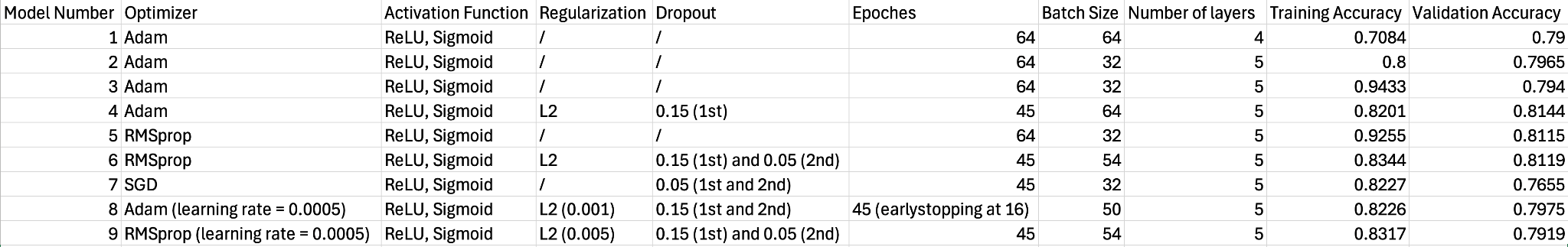
We realized that the uneven distribution of 1s and 0s was leading to inaccurate training results, so we implemented the following two approaches:

1. A scaler was used to apply equal weighting for features.
2. SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the number of 1s and 0s.

After using SMOTE, we made some progress in the training process. However, when running the first simple model, we found that we were facing an overfitting problem. To address this in subsequent training, we adjusted several aspects: changing the number of layers and the units in each layer, applying dropout, using L2 regularization, and modifying the learning rate to balance accuracy and overfitting.

While balancing accuracy and overfitting, we often encounter the issue of underfitting when using too much dropout. After comparing the training and testing graphs of each model, we selected the two best-performing results.

On the left, we used SGD as the optimizer, and on the right, we used RMSprop. We made significant improvements in addressing the overfitting issue. These improvements were achieved using dropout values of 0.15 and 0.05, applying BatchNormalization, and setting the learning rate to 0.005.



Before using dropout and L2 regularization, we compared the Adam and RMSprop optimizers. We observed that, under the same conditions, RMSprop produced better results than Adam. However, both models achieved training accuracy as high as 90%, and their test results were the best. Despite this, we cannot choose these models because the graphs show that their results could have been more stable and exhibited significant overfitting issues.

**Discussion, Conclusion, and Future Work**

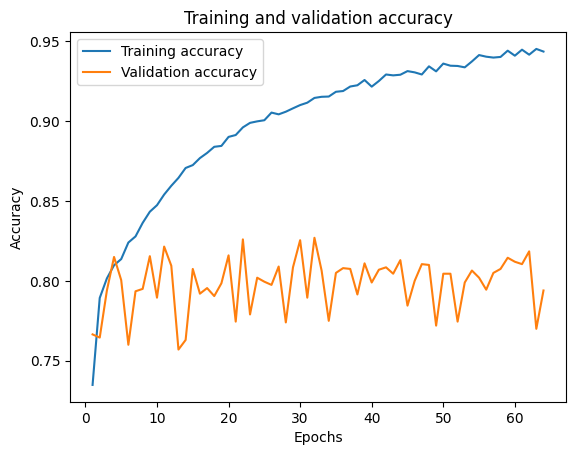
In this project, we aimed to predict customer churn for a bank using machine learning and deep learning. The main challenge was the imbalance between churned and non-churned customers, which often led to biased models. To tackle this, we used SMOTE to balance the data by generating synthetic churn cases. This helped our model generalize better.

Our model was based on a fully connected neural network, using ReLU activation and a sigmoid output. We trained it using optimizers like RMSprop and Adam. While we initially struggled with overfitting, adding dropout and L2 regularization improved accuracy to 86.42%, with an AUC score of 0.93. Although CNNs were tested, they didn’t perform significantly better than FCNNs.

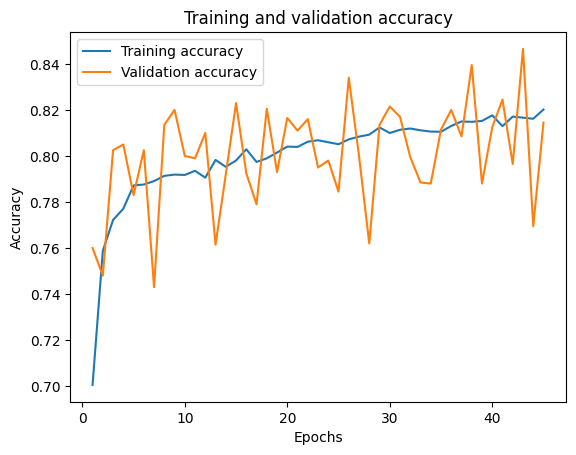
This project demonstrated the effectiveness of deep learning in predicting customer churn. By addressing class imbalance with SMOTE and fine-tuning our model, we achieved strong accuracy. However, overfitting remains a challenge that needs further research.

Future improvements include expanding the dataset with more customer data, exploring other architectures like RNNs, and creating real-time prediction systems. Additionally, using techniques like SHAP can enhance model explainability, helping banks better understand and act on churn predictions.

***Appendix***



Model: The model was more prone to overfitting without dropout and other regularization techniques like L2.



After implementing dropout and L2 regularization, the overfitting issue was significantly reduced; however, the results still exhibited considerable fluctuations.

***References***

Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

​​Huang, J., Chai, J. & Cho, S. Deep learning in finance and banking: A literature review and classification. *Front. Bus. Res. China* 14, 13 (2020). https://doi.org/10.1186/s11782-020-00082-6

Mujahid, M., Kına, E., Rustam, F. *et al.* Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *J Big Data* 11, 87 (2024). https://doi.org/10.1186/s40537-024-00943-4

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. *et al.* Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2, 1 (2015). https://doi.org/10.1186/s40537-014-0007-7

Suh, Y. Machine learning based customer churn prediction in home appliance rental business. *J Big Data* 10, 41 (2023). https://doi.org/10.1186/s40537-023-00721-8