MGSC – 695
PREDICTING CUSTOMER CHURN USING FEED-FORWARD NEURAL NETWORKS

Prepared for Fatih
Nayebi & Necmiye Gen
Yichen Yu
March 2024

I. Introduction

This report focuses on the utilization of feed-forward neural networks for predicting customer churn. In the contemporary business landscape, the ability to accurately forecast churn allows businesses to proactively address customer retention, directly influencing profitability and competitive advantage. The methodology in this report integrates advanced machine learning techniques, emphasizing the strategic configuration of neural network architectures, to refine the predictive accuracy of churn outcomes. This analysis aims to provide a comprehensive solution to churn prediction, underpinning its operational value with empirical evidence and offering insights into the practical applications of neural networks.

II. Dataset Overview

The dataset used in this analysis was sourced from Kaggle (https://www.kaggle.com/datasets/shubh0799/churn-modelling/data). In its original form, the dataset comprises 10,000 instances, each reflecting a unique customer profile, along with 14 distinct features.

(1). Data Preprocessing Steps

Pre-Processing Steps:

- **Irrelevant Variables Removal:** 'RowNumber', 'CustomerId', and 'Surname' were discarded to eliminate noise and focus on predictive features.
- **Final Feature Set:** Consolidated to 11 features to streamline the input for the neural network.
- **Target Variable:**
  - **'Exited':** Selected as the binary target variable to facilitate the model's classification task for churn prediction.
- **Feature Set:**
  - **Demographic Attributes:** 'Age' and 'Gender' provide insights into customer segments that may exhibit different churn behaviors.
- **Financial Metrics:**
  - 'CreditScore', 'Balance', and 'EstimatedSalary' are critical indicators of a customer's financial standing, which can influence churn.
  - **Customer Products:** 'NumOfProducts' reflects the depth of a customer's relationship with the business, affecting retention.
  - **Behavioral Indicators:** 'HasCrCard' and 'IsActiveMember' are included to gauge engagement and usage patterns, key predictors of churn.
- **Categorical Variables:**
  - **One-Hot Encoding:** Applied to 'Geography', 'Gender', and 'Tenure' to convert nominal categories into a numerical format that neural networks can process efficiently.

III. Experimental Setup

In the experimental setup, feature normalization was performed using the StandardScaler to ensure uniform scaling across all variables. The dataset was divided into training (60%), validation (20%), and testing (20%) sets to support a robust model evaluation. To be noted that the model development utilized PyTorch for neural network construction and training.

IV. Model 1 - Experiment with Simple Feed-Forward Neural Network Model

(1). Model Architecture

The Model 1 embodies a classic feed-forward neural network, architected with an input layer to accommodate 23 features, followed by two hidden layers packed with 64 neurons each. Meanwhile, the construction culminates in a singular output neuron that predicts customer churn. The network's blueprint intentionally lacks complex structures like dropout layers, relying instead on the robustness of fully connected layers. ReLU activation functions were strategically selected for their capacity to combat the vanishing gradient phenomenon, ensuring a stable gradient flow essential for deep learning models.

(2). Training Process

- **Learning Rate:** Set to 0.01 to ensure a balance between rapid convergence and stability during training.
- **Batch Size:** Determined to be 32, which optimizes computational efficiency and takes advantage of the stochastic nature of gradient descent for better model generalization.
- **Optimizer:** The Adam optimizer was selected for its ability to adjust learning rates dynamically, which is particularly advantageous in the context of high-dimensional datasets.
- **Regularization Strategy:** No explicit regularization techniques were applied for this model, suggesting this could be an area for future model improvement.

(3). Model Evaluation and Results

Model 1 demonstrated moderate success, with an accuracy of 79%, alongside precision and recall rates of 47% and 54%, respectively, culminating in an F1 score of 50% (Figure 1). These results, while indicative of the model's potential, also highlight areas that require further development. The training and validation loss graphs (Figure 2 and 3), alongside accuracy trends, signal a propensity for overfitting. This is evident from the rising validation loss, indicating that despite the model's aptitude in learning from the training data, it falls short on generalizing this knowledge to new, unseen data.

The elevated recall rate (Figure 1) showcases the model's improved detection capabilities for churn cases, but the less-than-optimal precision rate brings attention to a number of false positives—incorrectly predicted churn instances. As this model represents an exploratory step rather than a conclusive solution, it underscores the need for refinement. Future work will look into enhancing the model's architecture and potentially employing regularization techniques to

sharpen its predictive precision and to bolster its ability to generalize across new and unseen data.

V. Model 2 - Experiments with Different of Activation functions and Learning Rate with a Flexible Network Architectures

(1). Model 2 – Model Architecture

In the second stage of the pursuit for an improved model, the intricate process of parameter tuning is conducted for the neural network. The network's architecture remained constant with three hidden layers, implementing a descending neuron configuration from 128 to 64 and then 32 neurons, ensuring a funnel-like structure for feature consolidation.

For activation functions, the experiment was expanded to not only include the commonly used ReLU but also its variants—LeakyReLU, ELU, and PReLU. The rationale for these choices pivoted around optimizing the introduction of non-linearity, with a specific interest in how each function manages information flow, especially for negative input values. These functions were tested against one another to compare their individual impacts on model performance.

(2). Model 2 – Training Process

The training regimen for Model 2 included experimentation with learning rates. While keeping the initial rate of 0.01 as a baseline for rapid convergence, the slower rates such as 0.001 and 0.0001 were also investigated to discern the effects on the model's learning curve. The learning rate of 0.001 was ultimately selected for its efficiency in achieving a stable convergence, avoiding the pitfalls of excessive oscillations around the loss minimum.

The optimizer Adam was retained for its sophisticated mechanism of adjusting learning rates. Its computational prowess in high-dimensional spaces and ability to dynamically adapt learning rates to different parameters were instrumental for the nuanced training required in this stage. Notably, Model 2 did not yet incorporate explicit regularization methods as the focus was on assessing the base model's performance.

(3). Final Best Model Evaluation and Results

The optimal model was chosen for its ELU activation function and a learning rate of 0.001, standing out with an 82% accuracy and a macro average F1 score of 0.73. It's particularly adept at recognizing true churn cases, as shown by a recall rate of 66%, while maintaining a reasonable precision rate of 53% (Figure 4). This indicates that the model is not just predicting churn frequently but also with considerable accuracy, avoiding an excessive number of false positives. Such a balance is crucial for practical applications, ensuring that the model provides reliable predictions.

The graphical analyses from the model’s training—specifically the loss and accuracy plots (Figure 5 and 6), present a steady convergence of the model, with training and validation

losses displaying a consistent decline, and accuracy rates stabilizing after an initial phase of learning. This stability across epochs signifies a robust learning process without significant overfitting, an assertion that is corroborated by the validation loss tracing closely with the training loss. Moreover, the ROC curve analysis in Figure 7 indicates a well-performing model, with an area under the curve (AUC) of 0.84, suggesting a strong ability to differentiate between the positive class (customers who will churn) and the negative class (customers who will not churn).

An in-depth analysis of the results exposes that while the model's architecture is sound, the precision rate can still be enhanced. This would involve decreasing false positives to improve the model's overall reliability. Considering the results obtained, the model is proficient in its current form but has room for advancement, particularly in precision. Future iterations of the model might include a more extensive exploration of regularization methods or adjustments to the network architecture to further refine performance, especially in terms of generalizability to unseen data.

VI. Conclusion

Throughout the development of neural network models for churn prediction, challenges such as overfitting were aroused, where initial models performed well on training data but less so on unseen data. This was addressed by adjustments to the learning rate and the integration of different activation functions, which resulting find the final best model in the experiments. A model with an ELU activation function and a learning rate of 0.001, achieving an accuracy of 82% and a balanced precision-recall trade-off. However, the model's limitations point towards the potential for improvement. The precision rate, in particular, suggests a propensity for false positives that could be minimized with further tuning or advanced regularization techniques.

In conclusion, this exploration has yielded a powerful asset in the form of a churn prediction model, equipping businesses with actionable insights to retain customers more effectively. The findings underscore the potential benefits such predictive models hold for enhancing customer engagement strategies. Looking ahead, the model beckons further enhancements, inviting explorations into more sophisticated algorithms and a broader scope of data to fine-tune its predictions and embrace evolving customer trends.

VII. Appendix

Test Classification Report on Model 1:

	precision	recall	f1-score	support
Class 0	0.88	0.85	0.87	1607
Class 1	0.47	0.54	0.50	393
accuracy			0.79	2000
macro avg	0.67	0.70	0.68	2000
weighted avg	0.80	0.79	0.79	2000

Figure 1: Classification Report on Model 1

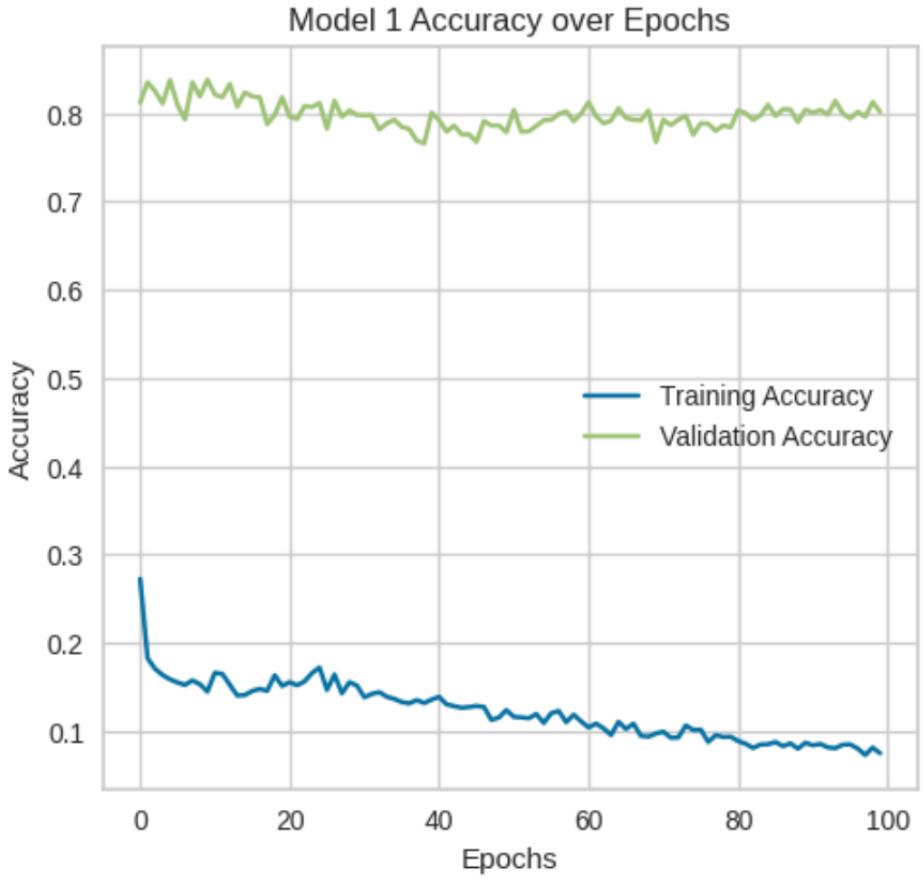

Figure 2: Model 1 – Model Accuracy Over Epochs

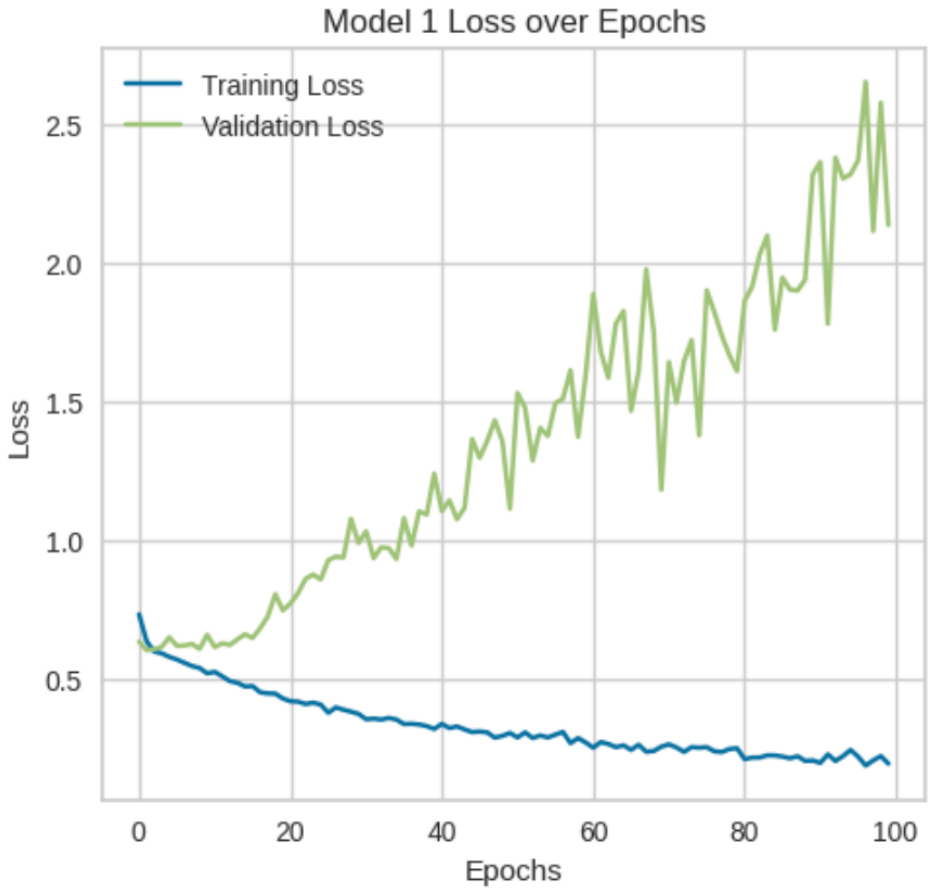

Figure 3: Model 1 – Model Loss Over Epochs

Classification Report on Best Model:

	precision	recall	f1-score	support
Not Exited	0.91	0.86	0.88	1607
Exited	0.53	0.64	0.58	393
accuracy			0.82	2000
macro avg	0.72	0.75	0.73	2000
weighted avg	0.83	0.82	0.82	2000

Figure 4: Classification on the Best Model

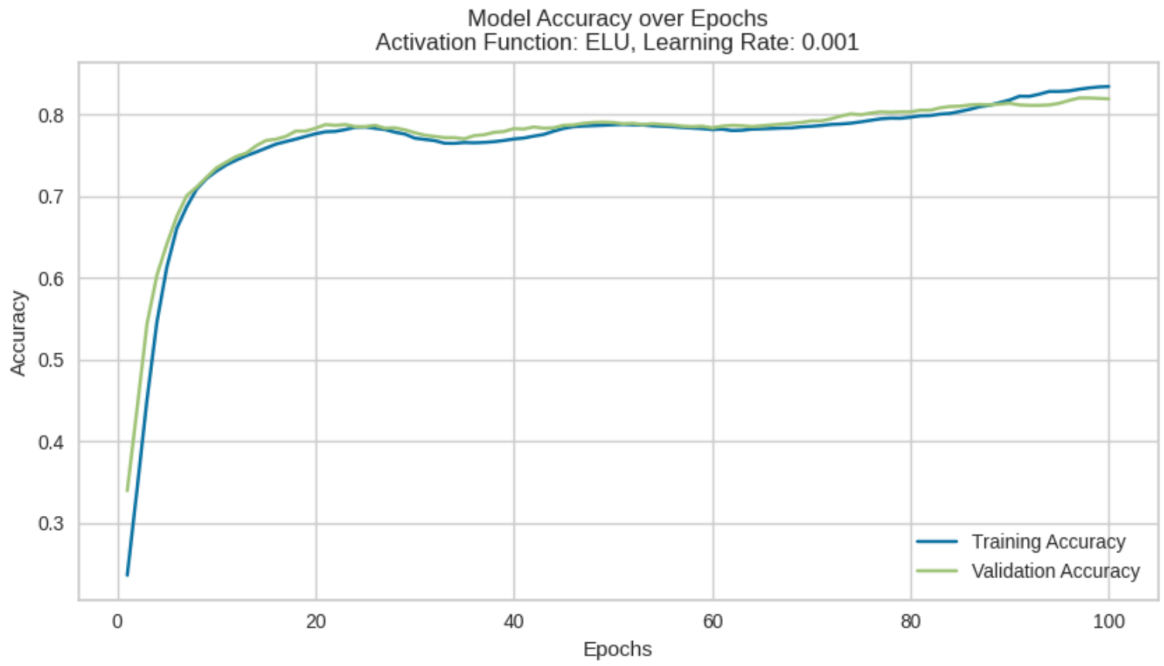

Figure 5: Best Model – Model Accuracy over Epochs

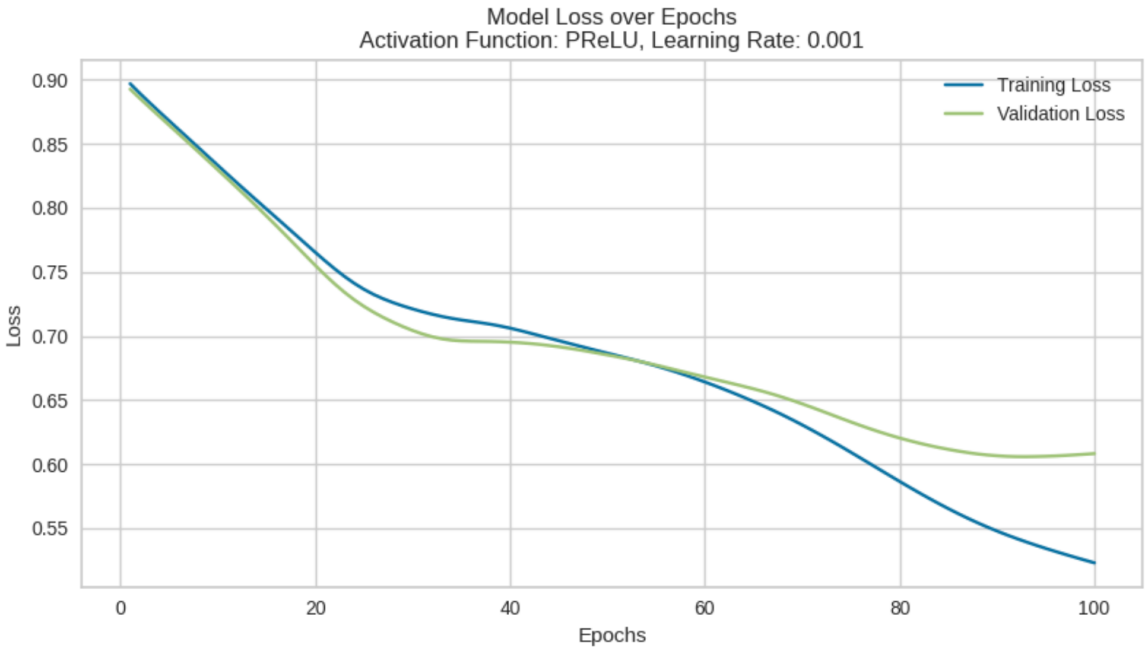

Figure 6: Best Model – Model Loss over Epochs

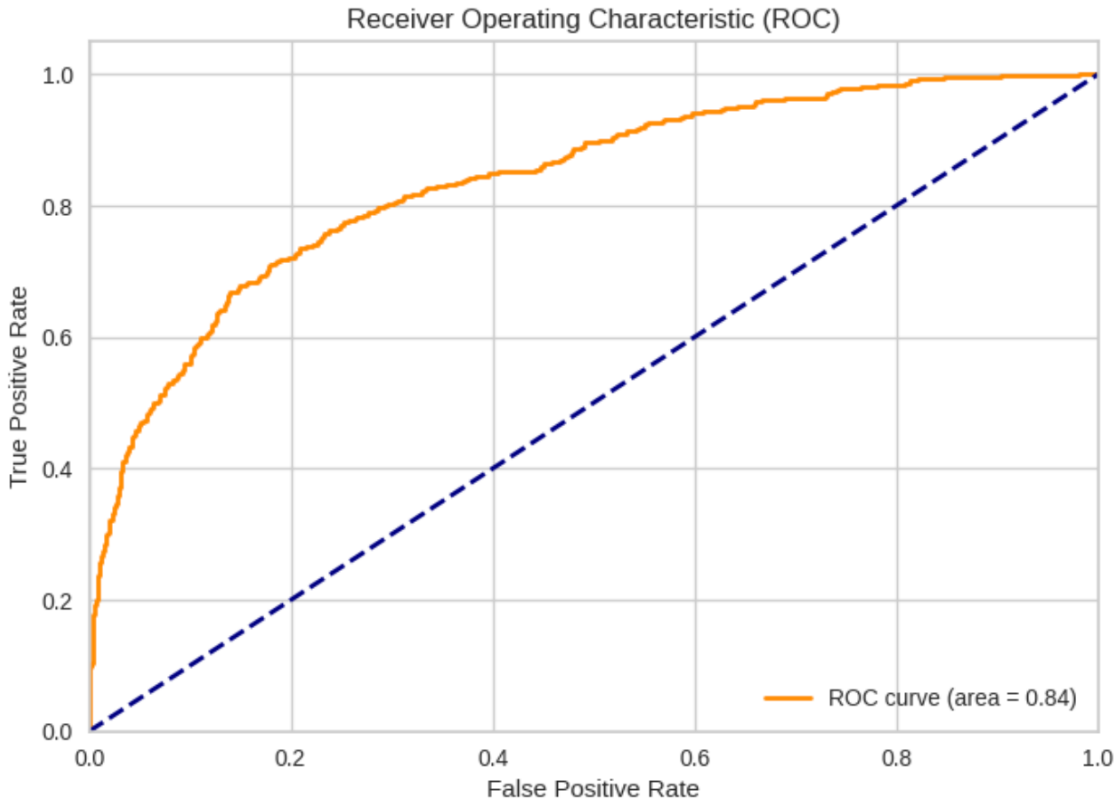

Figure 6: Best Model – ROC Curve