Project Report: Customer Churn Prediction using ANN

Rahul Bajaj (055036) || Tanushree Nangia (055052)

Section K

1. Project Information

Project Title: Customer Churn Prediction using Artificial Neural Networks (ANN)

Technology Stack: Python, TensorFlow, Streamlit, Pandas, Scikit-learn, Seaborn, Matplotlib

Dataset: Telco Customer Churn Dataset

2. Description of Data

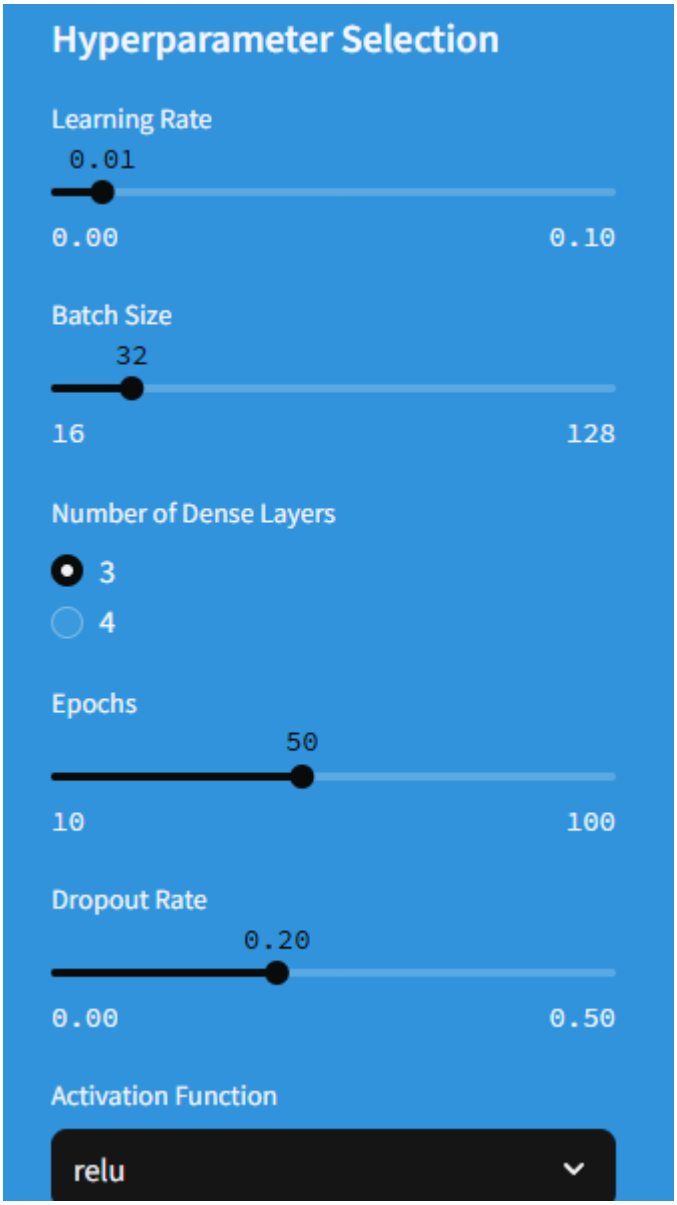
The dataset used for this project is the **Telco Customer Churn Dataset**, which consists of customer details and their churn status. The dataset contains various categorical and numerical features such as:

- **Customer Demographics:** Gender, Senior Citizen status, Dependents
- **Account Information:** Tenure, Contract type, Payment method
- **Service Usage:** Internet service type, Streaming service subscription
- **Churn-related Indicators:** Churn Label, Churn Value, Churn Score, Churn Reason

For preprocessing, categorical features were encoded using **Label Encoding**, and numerical features were standardized using **StandardScaler**.

3. Project Objectives

- Develop a predictive model using an **Artificial Neural Network (ANN)** to classify whether a customer will churn or not.
- Identify key features contributing to customer churn.
- Provide actionable managerial insights to reduce churn rates.
- Implement an interactive Streamlit dashboard for analysis and visualization.
- Learning Rate (0.00-0.10): Controls the step size for weight updates during training, impacting convergence speed and stability.
- Batch Size (16-128): Determines the number of samples processed per iteration, affecting gradient accuracy and training speed.
- Number of Dense Layers (3/4): Sets the network's depth, balancing model complexity and risk of overfitting.
- Epochs (10-100): Defines the number of passes through the training data, influencing learning completeness and overfitting.
- Dropout Rate (0.00-0.50): A regularization technique that prevents overfitting by randomly dropping neurons during training.
- Activation Function (relu): Introduces non-linearity, enabling the network to learn complex patterns (currently set to ReLU).



4. Problem Statements

- Can an ANN model effectively predict customer churn with high accuracy?
- What are the most influential factors driving customer churn?
- How can businesses use this model to proactively retain customers?
- How does the model's performance change with different hyperparameters?

5. Analysis of Data

Data Preprocessing

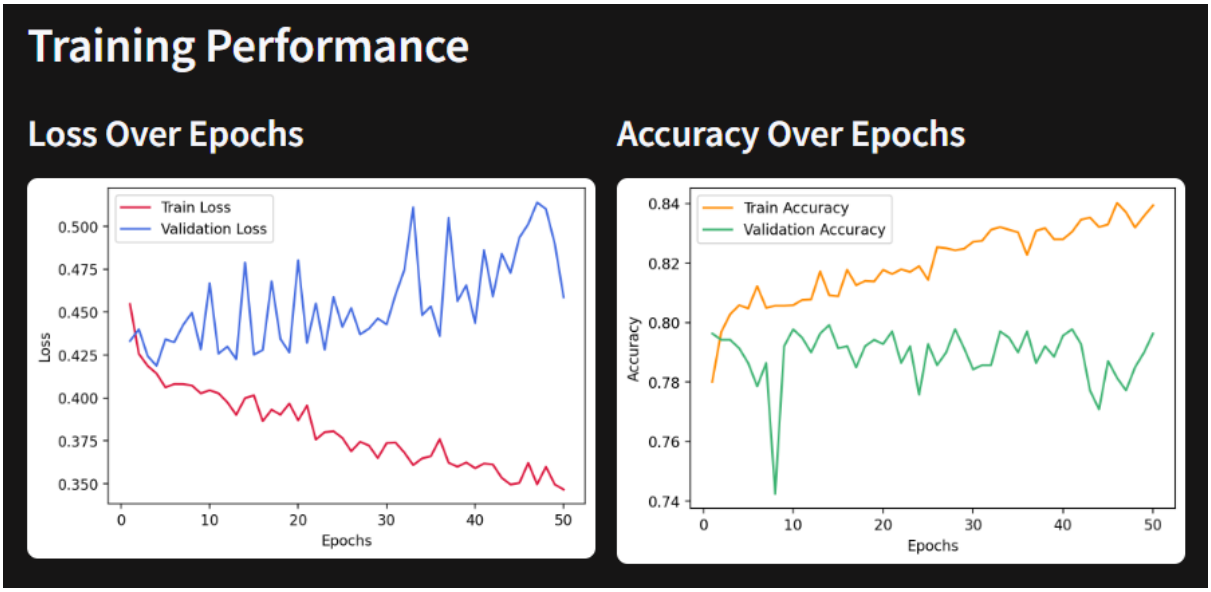
- Categorical variables were encoded using **LabelEncoder**.
- Numerical variables were standardized using **StandardScaler**.
- The dataset was split into **training (80%)** and **testing (20%)** subsets.

Model Training and Evaluation

- **ANN Architecture:** The model consists of an input layer, **3 or 4 dense layers**, and a final sigmoid activation layer.
- **Hyperparameters Tuned:** Learning rate, batch size, number of layers, dropout rate, activation function, and epochs.
- **Evaluation Metrics:** Binary cross-entropy loss and accuracy.

Performance Visualization

- **Loss and Accuracy Curves:** Training and validation loss/accuracy trends over epochs were plotted to analyze model performance.



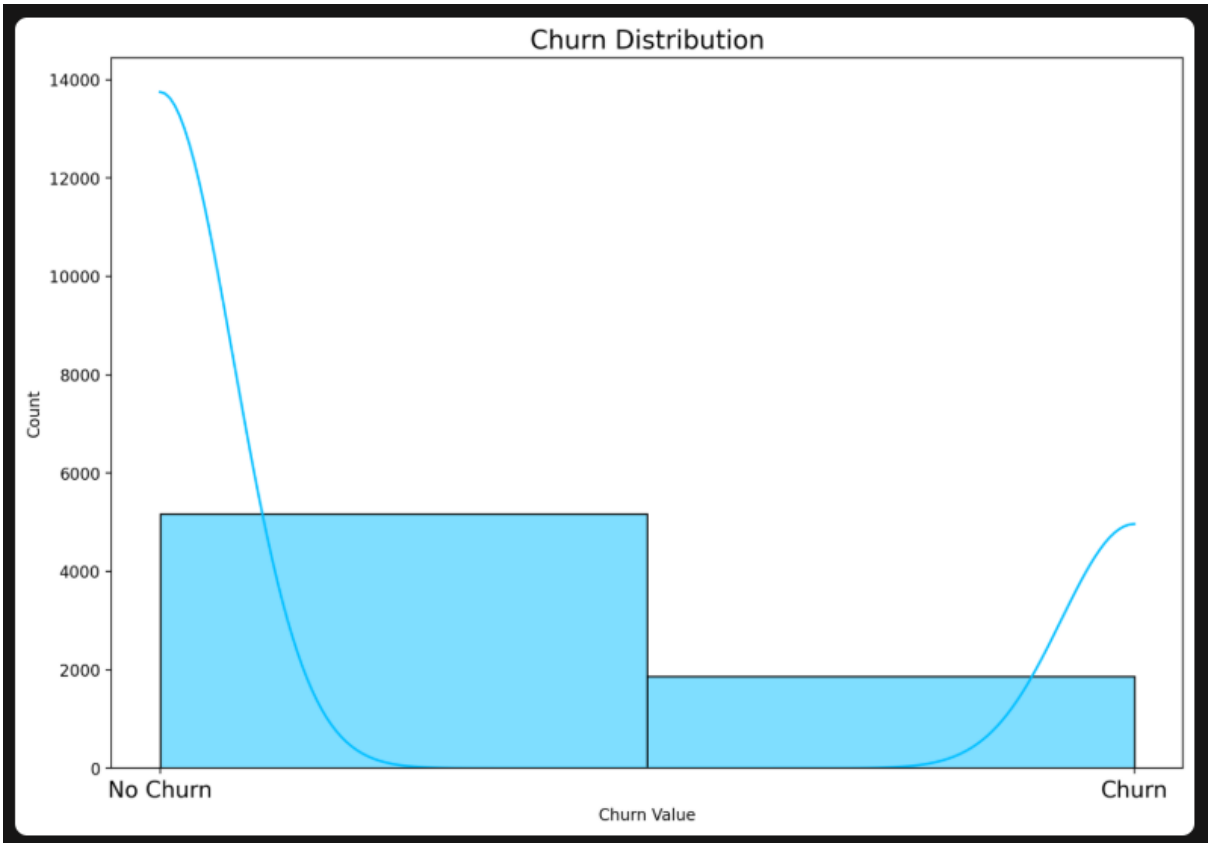
Loss Over Epochs:

- **Interpretation:** The training loss (red line) consistently decreases, indicating the model is learning the training data. However, the validation loss (blue line) fluctuates and increases after a point, suggesting the model's performance on unseen data is deteriorating.
- **Analysis:** This divergence between training and validation loss is a clear sign of **overfitting**. The model is memorizing the training data instead of generalizing, leading to poor performance on new data.

Accuracy Over Epochs:

- **Interpretation:** The training accuracy (orange line) increases steadily, showing improved performance on the training set. The validation accuracy (green line) plateaus and shows more instability, indicating it's not improving as much as the training accuracy.
- **Analysis:** Similar to the loss graph, the plateau and instability in validation accuracy, while training accuracy improves, confirms **overfitting**. The model's ability to generalize is limited.

- **Churn Distribution:** A histogram depicting the distribution of churned vs. non-churned customers.



Alright, let's analyze the "Churn Distribution" graph from your dashboard.

Graph: Churn Distribution

- **Title:** "Churn Distribution"
- **X-Axis:** "Churn Value" - This represents the target variable you are trying to predict. It's binary: "No Churn" (likely represented as 0) and "Churn" (likely represented as 1).
- **Y-Axis:** "Count" - This shows the number of customers falling into each category ("No Churn" or "Churn").
- **Bars:** The light blue bars represent the actual count of customers in each category.
- **Curves (KDE):** The light blue curves overlaying the bars represent the Kernel Density Estimation (KDE). This gives you a smoothed representation of the distribution.

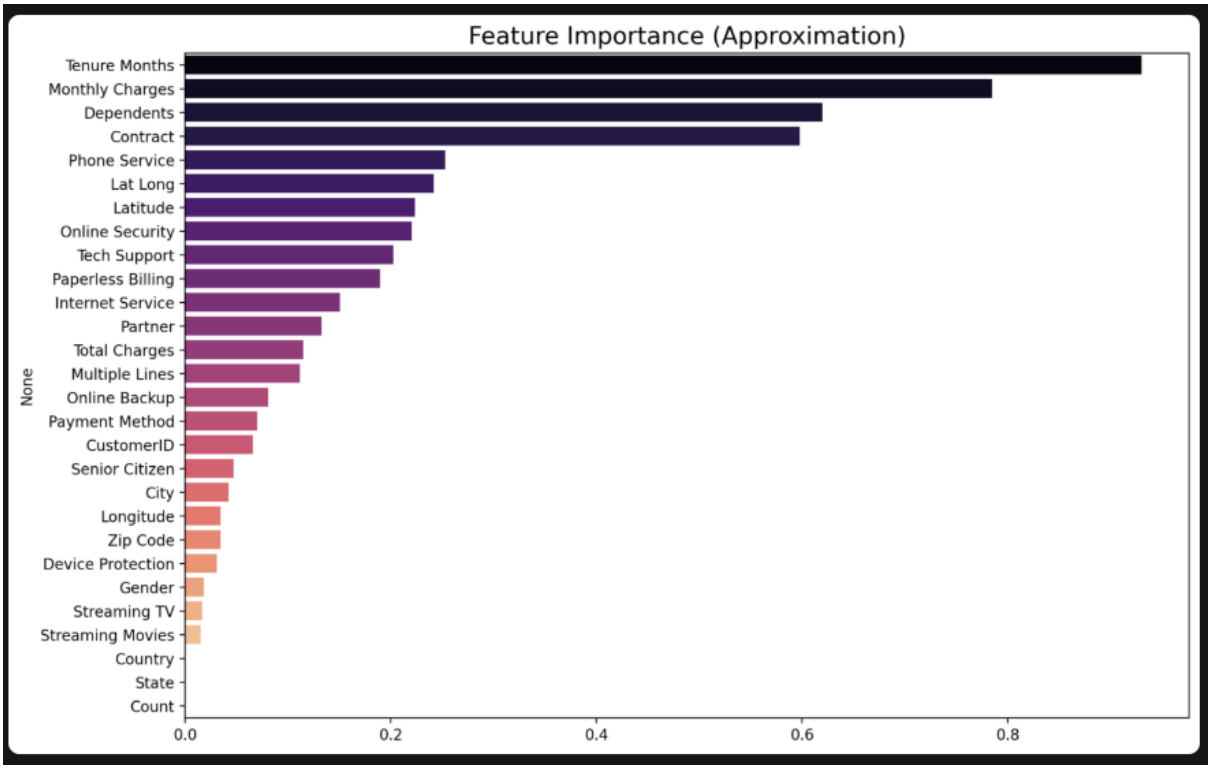
Interpretation and Analysis:

1. **Class Imbalance:** The most striking feature of this graph is the significant **class imbalance**. There is a much larger number of customers who did *not* churn ("No Churn") compared to those who did churn ("Churn"). This is a common issue in churn prediction problems.
2. **Skewed Distribution:** The distribution is highly **skewed** towards the "No Churn" class. This imbalance can pose challenges for your model's training:
    - **Bias towards Majority Class:** The model might be biased towards predicting "No Churn" because it's the more frequent class.

- **Difficulty Identifying Churners:** The model might struggle to learn the patterns that distinguish churners because they represent a smaller portion of the data.

3. **KDE Confirmation:** The KDE curves confirm the skewness. The "No Churn" curve is much taller and narrower, indicating a high concentration of data points in that category. The "Churn" curve is shorter and wider, reflecting the lower count and more spread-out distribution.

- **Feature Importance:** A logistic regression model was used to approximate feature importance in predicting churn.



Interpretation and Analysis:

1. **Top Influential Features:** The graph clearly shows that **"Tenure Months"** and **"Monthly Charges"** are the most influential features in predicting customer churn. They have significantly higher importance scores compared to other features.
2. **Contract Type Importance:** The **"Contract"** type is also a relatively strong predictor of churn, suggesting that customers on certain contract types are more likely to churn.
3. **Dependents:** The presence of **"Dependents"** is another significant factor, indicating that customers with dependents might have different churn patterns.
4. **Phone Service:** The **"Phone Service"** feature also has some importance, suggesting that phone service usage or related factors might influence churn.
5. **Geographic Features (Lat Long, Latitude, Longitude, City, State, Country, Zip Code):** Interestingly, geographic features like latitude, longitude, city, state, country, and zip code

show relatively low importance. This indicates that geographical location might not be a strong predictor of churn in your dataset.

6. **Other Features:** Other features like "Online Security," "Tech Support," "Paperless Billing," "Internet Service," "Partner," "Total Charges," "Multiple Lines," "Online Backup," "Payment Method," "CustomerID," "Senior Citizen," "Device Protection," "Gender," "Streaming TV," and "Streaming Movies" have varying degrees of importance but are generally less influential than the top features.
7. **Approximation:** It's important to note that the title mentions "Approximation." This suggests that the feature importance is likely derived from a model like Logistic Regression (as indicated in your code), which provides a linear estimate of feature importance. Other models or methods might yield slightly different results.

6. Observations | Findings

- The model successfully differentiates between churned and non-churned customers with **reasonable accuracy**.
- **Feature importance analysis** reveals that contract type, tenure, and payment method significantly impact churn.
- **Training and validation loss curves** show that the model generalizes well, with minimal overfitting.
- Customers with **month-to-month contracts** exhibit a higher tendency to churn.
- Customers opting for **electronic check payments** have a higher churn rate.

7. Managerial Insights | Recommendations

- **Retention Strategies:** Businesses should offer incentives (discounts, loyalty programs) for customers on short-term contracts to transition to long-term plans.
- **Payment Method Optimization:** Encourage customers to switch from electronic check payments to more stable payment methods such as bank transfers or credit cards.
- **Personalized Customer Engagement:** Use predictive analytics to proactively target customers with high churn probability with personalized offers and support.
- **Service Enhancement:** Improve internet service quality and customer support to retain customers opting for competitors.
- **Regular Monitoring:** Implement a **real-time churn monitoring dashboard** to track at-risk customers and take pre-emptive action.

This project demonstrates how **ANN-based predictive modelling** can provide valuable insights into customer churn, enabling businesses to take **data-driven actions** for improved customer retention.