Student Name: Aniruddha Joshi

Data Set: <https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset/data>

**Milestone 3**

**Project Topic:** Predicting Customer Retention Based on Subscription Model

**Business Problem:** This project aims to predict customer churn for a subscription-based service and identify key factors contributing to churn. By predicting which customers are likely to churn, the company can implement targeted retention strategies to minimize revenue loss. Customer churn occurs when customers end their relationship or subscription with a company. It is a key metric that affects a company's revenue, growth, and customer retention. High churn rates can indicate issues such as poor customer service, product dissatisfaction, high competition, incorrect pricing or ineffective marketing strategies. Addressing churn proactively can improve customer retention rates, maximize customer lifetime value, and ensure long-term sustainable growth.

**Background/History**

The project is focused on customer churn within a subscription-based service model. This involves understanding the importance of customer retention for long-term business success. The available data includes historical customer information, usage patterns, demographics, contract details, and churn status. The project seeks to leverage this data to build predictive models that can help the business understand and address the problem of customer churn.

**Data Explanation**

Data Source: The primary dataset is a customer churn dataset from Kaggle.

Link is: <https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset/data>

Data Format: The data is provided in a CSV file format.

Data Contents: The dataset contains up to 12 variables.

These include:

CustomerID, Age, Gender, Tenure, Usage, Usage Frequency, Support Requests, Support Calls, Payment Delay, Subscription Type, Contract Length, Total Spend, Last Interaction, Churn

**Data Preparation:**

Libraries: The project uses Python libraries such as Pandas, NumPy, and Seaborn for data cleaning, manipulation, and visualization.

Cleaning: The data is cleaned by removing invalid or incomplete rows and filling missing data with mean or averages. The data types of columns were converted to appropriate types.

Categorical Data: Categorical data is converted using one-hot encoding as required. The 'Gender', 'Subscription Type' and 'Contract Length' columns were converted using label encoding.

Data Exploration: The data was explored using functions to find missing data, outliers, maximum and minimum values.

Scaling: Numerical features were scaled using StandardScaler.

Data Dictionary: The data dictionary is provided in the "Data Contents" section above. It describes the type of information contained in each column.

**Methods:**

Data Exploration:

Pandas, NumPy and Seaborn libraries were used to clean the data by removing invalid/incomplete rows and filling missing data.

Visualizations such as scatter plots, histograms and box plots were created using Matplotlib, Plotly and/or Seaborn to identify any relationships between variables.

A screenshot of a graph

AI-generated content may be incorrect.

Figure Correlation Matrix of customer usage

A grid of blue squares

AI-generated content may be incorrect.

Figure Correlation Matrix of variables

Data distributions, outliers, and missing values were explored.

**Feature Engineering**:

Feature engineering may not be required with the given variables. A cross-reference matrix review of the variables was done to ensure to identify the relationship.

**Machine Learning Modeling:**

Following models were considered: Logistic Regression, Random Forest, SVM, Neural Networks. The best model was to be evaluated. For neural networks the Scikit-learn libraries were used.

The data was split into training and testing sets, with 80% used for training and 20% for testing.

The Neural Network model was defined with input layers, hidden layers (with ReLU activation), and an output layer (sigmoid activation).

**Model Evaluation:**

Model performance was evaluated using metrics like accuracy and loss. The accuracy, precision, recall, and f1-score were calculated for various model. Below are the results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Accuracy | Precision (False/True) | Recall (False/True) | F1-Score (False/True) |
| Logistic Regression | 0.8511 | 0.81 / 0.88 | 0.85 / 0.85 | 0.83 / 0.87 |
| Random Forest | 0.9996 | 1.00 / 1.00 | 1.00 / 1.00 | 1.00 / 1.00 |
| SVM | 0.9793 | 0.96 / 1.00 | 1.00 / 0.97 | 0.98 / 0.98 |
| Scikit-learn NN | 0.9995 | 1.00 / 1.00 | 1.00 / 1.00 | 1.00 / 1.00 |
| Keras NN | 0.9993 | 1.00 / 1.00 | 1.00 / 1.00 | 1.00 / 1.00 |

**Analysis**

Data Visualization: Data visualization was used to find relationships between the variables in the dataset. A correlation matrix visualized correlations between numerical features. Pair plots were used to see distributions and relationships. Box plots were used to visualize the distribution of numerical variables across different categories.

A blue green orange and orange shapes

AI-generated content may be incorrect.

Figure Violin Plot of Total Expenditure

A diagram of a chart

AI-generated content may be incorrect.

Figure Box plot to find outliers in Subscription Types

**Model Performance:** The Scikit-learn NN model achieved high accuracy, precision, recall and f1-score amongst other models evaluated.

**Conclusion**

The project successfully developed a model to predict customer churn using a Scikit-learn NN network. The model's high accuracy suggests that it can be used to identify customers at risk of churning. This allows the business to take proactive measures to retain these customers.

**Assumptions**

The assumption is the provided dataset is representative of the overall customer base and historical data patterns are indicative of future trends.

**Limitations**

Feature Engineering: Although feature engineering may not have been required, it may be that better features could be engineered to improve the model.

Data Completeness: The data may not be representative of all customer base.

**Challenges**

Data quality and completeness presented challenges. Feature engineering was considered and found to be unnecessary. Testing and finding the accuracy of different models was a challenge. Each model had different setting requirements and tuning imposed challenges.

**Future Uses/Additional Applications**

The model can be used to identify customers at risk of churning in real-time. The model can be used to develop targeted marketing campaigns for customer retention. Further analysis can be performed to find the drivers of customer churn by looking at feature importances from the models. The model can be extended by including new features that might improve the model.

**Recommendations**

Implement the model in real-time to predict customer churn. Use model insights to develop targeted retention strategies. Continuously monitor and evaluate the model's performance. Explore additional data sources and features to improve the model's accuracy.

**Implementation Plan**

Develop a data pipeline to collect and prepare data for model input. Develop continues flow of data. Integrate the model into the company's existing systems and order management software. Train the model with new data periodically to ensure accuracy and then deploy the model in the production environment. Monitor model performance and make improvements as necessary.

**Ethical Assessment**

Privacy Concerns: All data used must be anonymized to protect customer identities.

Gender Bias: Model fairness was assessed to ensure no discriminatory bias against specific genders.

**References:**

1. (Kokkula, n.d.; Azeem, 2023) Customer churn dataset
2. Kaggle dataset link: [https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset/data](https://www.google.com/url?sa=E&q=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fmuhammadshahidazeem%2Fcustomer-churn-dataset%2Fdata)