Disha Devalia (devaliadisha06@gmail.com)  
  
**Customer Churn Prediction Project Documentation**

1. **Introduction**

At Sunbase, our unwavering commitment to customer satisfaction is the driving force behind our mission. We recognize the paramount importance of not only acquiring new customers but also nurturing our existing relationships. In line with this dedication to our valued customers, we have embarked on a data-driven initiative aimed at predicting customer churn.

As a Machine Learning Intern here at Sunbase, you have been entrusted with the pivotal responsibility of harnessing the capabilities of machine learning to construct a robust predictive model for identifying potential customer churn. This project stands as an integral component of our mission, empowering us to proactively identify potential churn, and in turn, implement strategic measures that will elevate customer satisfaction.

1. **Project Objective**

The central goal of this project is to craft a machine learning model capable of precisely forecasting customer churn. By delving into our reservoir of historical customer data, our aim is to unearth invaluable insights into customer behavior and the underlying factors that contribute to churn. The predictive model we construct will enable Sunbase to take pre-emptive actions in retaining our customers, reinforcing our customer relationships, and nurturing the enduring prosperity of our business.

1. **Data Source**

In support of this project, we have supplied you with a CSV-format dataset. This dataset is a treasure trove of historical customer information, encompassing an extensive array of attributes, interactions, and, most crucially, the binary indicator of whether a customer has churned or not. This dataset serves as the bedrock upon which you will lay the foundation for your machine learning model.

1. **Project Roadmap**

Your voyage as a Machine Learning Intern in this project will traverse a comprehensive pipeline. This journey encompasses data preprocessing, the craft of feature engineering, the selection and training of machine learning models, meticulous model evaluation, and the ultimate deployment of the chosen model. This roadmap equips you with the power to make decisions rooted in data, provide insights that hold significance, and contribute to Sunbase's resolute customer-centric strategy.

This document is your canvas to depict each step of this remarkable project. It is an opportunity to unveil the methodologies employed, record the outcomes, and extract key insights that will steer our course. We extend to you an invitation to embark on this exciting expedition, one that will ultimately redefine how we, at Sunbase, approach customer satisfaction.

Let us begin our journey into the depths of this project.

**Table of Contents**

1. **Project Overview**
2. **Data Preprocessing**
3. **Exploratory Data Analysis (EDA)**
4. **Feature Engineering**
5. **Model Development**
6. **Model Optimization**
7. **Model Deployment**
8. **Dashboarding (Tableau Dashboard) and Reporting**
9. **Documentation and Reporting**
10. **Conclusion and Recommendations**
11. **References**

**1. Project Overview**

At Sunbase, our unwavering commitment revolves around the satisfaction of our valued customers. I understand that not only acquiring new customers but also retaining our existing ones is pivotal to our mission. In alignment with this dedication, I have embarked on a data-driven initiative with the primary goal of predicting customer churn.

**1.1 Objective**

The central objective of this project is to harness the potential of machine learning to construct a robust model for forecasting customer churn. As a Machine Learning Intern at Sunbase, I am tasked with spearheading this endeavor, which holds immense significance. By doing so, I can proactively identify potential churn events and take strategic actions to enhance customer satisfaction.

**1.2 Importance**

This project bears critical importance within our organizational mission. It allows us to gain deeper insights into customer behavior and the factors influencing churn. Moreover, the predictive model I aim to develop will empower Sunbase to take preemptive measures in retaining customers, ultimately fortifying our customer relationships and contributing to the long-term prosperity of our business.

**1.3 Expected Outcomes**

By the successful implementation of this project, I anticipate several key outcomes:

* Improved Customer Retention: With the ability to predict churn, I can take proactive steps to retain customers and reduce churn rates.
* Enhanced Customer Satisfaction: Customers are more likely to stay when their needs and concerns are addressed before they churn.
* Data-Driven Decision-Making: The model's insights will guide my decision-making processes and foster a customer-centric approach.
* Sustainable Business Growth: The project contributes to our long-term growth and stability by retaining and satisfying our customer base.

Join me on this exciting journey to develop a predictive model that will revolutionize how we approach customer satisfaction at Sunbase. In the following sections, I will delve into the intricate details of the project, commencing with the crucial data preprocessing phase.

**2. Dataset and Data Preprocessing**

In this section, I will outline the essential steps taken in the data preprocessing phase, which is a critical foundation for the subsequent stages of the project.

* 1. **Dataset**
* The dataset is loaded from an Excel file named "customer\_churn\_large\_dataset.xlsx." It contains historical customer information.
* It includes:
  + CustomerID: A unique identifier for each customer.
* Name: The customer's name.
* Age: Age of the customer.
* Gender: Customer's gender (e.g., Male, Female, or other).
* Location: The customer's location or place of residence.
* Subscription\_Length\_Months: The duration of the customer's subscription in months.
* Monthly\_Bill: The amount of money the customer is billed on a monthly basis.
* Total\_Usage\_GB: The total data or usage in gigabytes (GB) by the customer.

**2.2 Data Cleaning**

Data quality is of utmost importance, and thorough data cleaning was conducted to ensure the dataset's integrity. The key activities in this phase include:

* Identifying and addressing outliers and anomalies, which is essential for maintaining the accuracy and reliability of our data.
* Removal of duplicate records to prevent any skewing of results and ensure unbiased analysis.
* Addressing potential data entry issues to maintain the overall quality and consistency of the dataset.

**2.3 Handling Missing Values and Outliers**

Dealing with missing data is a common challenge in data analysis. A meticulous approach was taken to manage missing values:

* Various strategies were employed, including imputation, to address missing data points and preserve the completeness of the dataset.
* A careful evaluation of the impact of these strategies on the analysis was conducted to ensure the robustness and validity of the results.
* Various data visualization techniques, including box plots, were utilized to visualize the distribution of numerical features such as "Age," "Subscription\_Length\_Months," "Monthly\_Bill," and "Total\_Usage\_GB." These visualizations helped identify potential outliers and understand the data's statistical properties.
* A blue squares with white lines

  Description automatically generated
* A blue rectangular object with black lines

  Description automatically generated
* A blue squares with white lines

  Description automatically generated with medium confidence

**2.4 Preparing Data for Machine Learning:**

* Encode categorical variables into a numerical format, create training and testing sets for model development and evaluation.

A screenshot of a computer

Description automatically generated

* Gender Encoding: The "Gender" variable is encoded using one-hot encoding. This process creates binary columns for each category, typically with 0s and 1s, where 1 indicates the presence of a category and 0 indicates its absence. This transformation allows the machine learning model to understand and work with gender information effectively.
* Location Encoding: The "Location" variable is also encoded using one-hot encoding. Each unique location category is transformed into a binary column, making it easier for the model to handle location-based information.
* Train-Test Split:

A common method for creating the training and testing sets is by using the train\_test\_split function. This function randomly shuffles the dataset and divides it into the training and testing sets while maintaining the distribution of the target variable (in this case, "Churn"). The split ratio (e.g., 80% training and 20% testing) is determined based on the project's requirements.

**3. Feature Engineering:**

Feature engineering is a pivotal step in constructing a predictive model. In this phase:

**3.1 Generate Relevant Features:**

* Create new features or transform existing ones to capture valuable insights related to customer churn prediction. These features should be selected based on their potential to improve the model's predictive accuracy.
* Example:
* TenureInDays: Created to represent the customer's tenure in days, offering a more granular view of customer loyalty.
* TotalMonthlyCost: Calculated as the product of "Monthly\_Bill" and "Subscription\_Length\_Months," reflecting the total cost incurred by customers.
* UsageToCostRatio: Established by dividing "Total\_Usage\_GB" by "TotalMonthlyCost," providing insights into the relationship between usage and costs.
* Location-Specific Churn Rate: Generated location-specific churn rates by averaging churn for each unique location category, allowing geographic analysis of churn patterns.
  1. **Apply Feature Scaling or Normalization**:

Min-Max scaling was applied to several numerical features to normalize their values within a consistent range. This scaling was performed on the following features:

Age: Scaled to bring age values within a consistent range.  
TenureInDays: Ensured uniformity in tenure values by scaling.  
Monthly\_Bill: Transformed to a standardized scale to maintain consistency across customers.  
Total\_Usage\_GB: Normalized to make usage data comparable across customers.  
TotalMonthlyCost: Scaled to a consistent range to facilitate model training.

These data preprocessing steps were meticulously executed to prepare the dataset for model development and evaluation. They lay the groundwork for ensuring that the data used for machine learning analysis is dependable, accurate, and optimized for predicting customer churn.  
Join me as we progress further into the project, unveiling more insights and the model development process.

**4. Exploratory Data Analysis (EDA)**

* 1. **Class Distribution Analysis:** The class distribution of the target variable "Churn" was analyzed to assess data balance. This step helps in understanding whether there is an imbalance in the dataset, which could affect model training.

A blue rectangular bars with white text

Description automatically generated

* 1. **Feature Correlation Analysis:**   
     A correlation matrix was generated to evaluate the relationships between features and the target variable "Churn." This analysis helped identify which features were most correlated with customer churn.

A screenshot of a graph

Description automatically generated

* 1. **Feature Importance Analysis:**   
     A machine learning model (XGBoost) was used to determine feature importances, providing insights into which features play a crucial role in predicting customer churn.

A screen shot of a computer

Description automatically generated

**5. Model Building:**

* 1. **Choosing Appropriate Machine Learning Algorithms:**

The project considered multiple machine learning algorithms, including but not limited to logistic regression, random forest, and neural networks. These algorithms were selected based on their suitability for the churn prediction task. Logistic regression is a fundamental algorithm for binary classification, random forests offer ensemble learning capabilities, and neural networks can capture complex patterns in data.

* 1. **Training and Validation:**

The selected machine learning model, or models, were trained on the training dataset. During the training process, the model learns from historical customer data to understand patterns that might lead to churn. Cross-validation techniques were used to validate the model's performance. Cross-validation helps in assessing how well the model generalizes to unseen data by splitting the training data into multiple subsets for training and validation.

* 1. **Evaluation of Model Performance:**

The model's performance was evaluated using a set of relevant metrics. These metrics include:

* + Accuracy: Measures the proportion of correctly predicted customer churn and non-churn cases.
  + Precision: Focuses on the fraction of true churn predictions out of all predicted churn cases, indicating the model's ability to make accurate positive predictions.
  + Recall (Sensitivity): Assesses the ability of the model to correctly identify churn cases out of all actual churn cases.
  + F1-Score: A balance between precision and recall, offering a combined measure of the model's accuracy in predicting both positive and negative classes.

These evaluation metrics provide a comprehensive assessment of the model's effectiveness in predicting customer churn and guide the selection of the best-performing model for deployment. The goal is to achieve a balance between precision and recall while maximizing overall accuracy.

**6. Model Optimization:**  
The primary goal is to fine-tune the chosen machine learning model to improve its predictive performance. This phase involves iterative experimentation and adjustment of various model parameters and hyperparameters.

* 1. **Hyperparameter Tuning:**

Various hyperparameters specific to the selected machine learning models (Logistic Regression, Random Forest, Neural Network) were fine-tuned.

For instance, in the case of Logistic Regression, hyperparameters like 'C' and 'penalty' were optimized to achieve better performance.

Logistic regression and Random Forest:

A computer screen shot of a number

Description automatically generated

Neural network:

A black screen with white text

Description automatically generated

* 1. **Cross-Validation:**

Cross-validation techniques were employed to assess the model's performance across different subsets of the training data.

This helped in selecting the best combination of hyperparameters and ensured the model's generalization ability.

* 1. **Feature Selection and Ensemble Methods:**

Feature selection techniques were applied to identify the most relevant features for predicting customer churn. By focusing on key features, the efficiency and predictive accuracy of the models were enhanced.  
  
Ensemble methods, such as Random Forest, were considered as part of the model optimization process. These methods were used to combine multiple models and create a more powerful predictive model.

* 1. **Regularization and Model Comparison:**

In the case of neural networks, regularization techniques (e.g., L1 and L2 regularization) were explored to prevent overfitting and improve the generalization ability of the model.  
  
The performance of different model iterations was compared, and various performance metrics (e.g., accuracy, precision, recall, and F1-score) were evaluated on validation datasets. This comparison helped in selecting the final model configuration. Neural Network proved to be the best\_model.  
A graph with lines and numbers

Description automatically generated

**Neural Network proved to be the best-model out of all and was selected for prediction.**

**7. Model Deployment:**

It involves making the model available for predictions or inferences based on new, unseen data.   
Model deployment is the bridge between the model development phase, where the model is created and tested, and its practical application in solving real-world problems.  
  
 **7.1** **Handling New Customer Data**:

Ensuring that the deployed machine learning model is capable of handling new customer data, which allows the model to take customer information, such as age, subscription length, monthly bill, and total data usage, and provide churn predictions for individual customers.   
New data is used as input to the model, and it generates predictions in real-time or near real-time.

**7.2 Web Hosting and Flask**:   
To make the model accessible to our teams and systems, Flask was chosen for deploying the machine learning model as a web service.  
Flask provides a web interface or API where users can input customer data, and the model can provide churn predictions.   
This interface is efficient in handling requests and responses, ensuring scalability as our model's usage grows.

A screenshot of a computer

Description automatically generated

By default, Flask's development server runs on port 5000.  
  
Flask was selected for several reasons:

* **Lightweight**: Flask is minimalistic and easy to set up, making it suitable for our project's size and requirements.
* **Scalability**: It efficiently handles requests and responses, enabling us to scale the model's usage as needed.
* **Integration**: Flask seamlessly integrates with other Python libraries and frameworks, making it versatile for deploying machine learning models.

**7.3 Churn Predictions**:

The code ensures that the model's predictions are based on the features and data patterns it has learned during its training phase. The model has been trained to recognize patterns associated with customer churn.  
  
To provide churn predictions, the code takes features like age, subscription length, monthly bill, and total data usage as input. These features are crucial for making accurate predictions.  
  
The model generates binary predictions: 1 for customers likely to churn and 0 for those likely to stay. This binary approach simplifies the churn prediction process and allows for actionable insights.

A screenshot of a computer

Description automatically generated

**8. Dashboarding and Reporting**A Tableau dashboard was created to provide an overview of key insights and visualizations derived from the project.

A screenshot of a data

Description automatically generated

* **Churn Rate by Location Heatmap:**
  + A heatmap visualizing churn rates across various geographic locations, helping identify regions with higher churn percentages.
* **Distribution Category Pie Chart:**
  + A pie chart representing the distribution of customers across different segments or categories, offering insights into your customer base's composition.
* **Total Number of Customers Value:**
  + A numeric value showcasing the overall count of customers in the dataset.
* **Total Churned Value:**
  + A numeric value indicating the total number of customers who have churned.
* **Churn Rate Value:**
  + A numeric value expressed as a percentage, signifying the portion of customers who have churned relative to the total customer count.
* **Total Monthly Cost for Churned Customers by Location Bar in Bar:**
  + A bar-in-bar chart illustrating monthly costs incurred by churned customers, segmented by different geographic locations, highlighting cost trends.
* **Subscription Length vs. Churn Rate Comparison Bar Chart:**
  + A bar chart comparing churn rates based on different subscription lengths, providing insights into how subscription duration impacts customer retention.

**9. Documentation and Reporting:**

**9.1 Code Comments**  
Throughout the project, code comments were added to provide explanations and context for different parts of the code.   
Code comments are essential for code maintainability and readability.   
They explain the logic and purpose of functions, classes, and sections of code.

**9.2 Reporting Formats**Different reporting formats were employed to convey project progress, findings, and results. This may include written reports, and visual dashboards.  
For example, Tableau dashboards provide visual and interactive reports, python visualization provide insights, while written documents offer more detailed explanations.

**10. Conclusion and Recommendations:**  
In conclusion, our project created an effective customer churn prediction model that empowers Sunbase to proactively address churn, strengthen customer relationships, and drive business growth.   
Key findings include regional churn patterns and the impact of subscription length.   
We recommend focusing on high-churn regions and continuous model monitoring and retraining for ongoing success in improving customer satisfaction and retention.

1. **References:**

In the references section, due credit is given to the sources of data, libraries, frameworks, or algorithms that played a vital role in the successful execution of this project.   
This includes acknowledging external resources, research papers, or datasets that contributed to the development and implementation of our customer churn prediction model.

Top of Form