**Project Plan**

**Name: Blessing Nkoyeasua Iwerumor**

**Student No: 22069644**

**Introduction:**

This project plan sets the stage for exploration of customer churn prediction, underpinned by a meticulously crafted timeline and robust management strategy, aiming to distill algorithmic efficacy into actionable insights for businesses.

**1. Project Title:** Comparative Analysis of Machine Learning Algorithms for Customer Churn Prediction

**2. Research Question:**

1. Which machine learning algorithm -Stochastic Gradient Boosting (SGB), Support Vector Machine (SVM) and Random Forest provides the most accurate predictions for customer churn?
2. How can using hyperparameter tuning techniques like Grid Search or Random Search enhance the performance of the algorithms in predicting customer churn?

**3. Project Objectives**

1. To identify and compare the performance of various machine learning algorithms for predicting customer churn.
2. To evaluate the accuracy, precision, recall, and F1-score of each algorithm.
3. To recommend the most effective algorithm for customer churn prediction based on the analysis.
4. To employ Random Search or Grid Search techniques to optimize the hyperparameters of the models, with the aim of enhancing their performance in predicting customer churn.

**Background**

Customer churn, the phenomenon where customers stop using a company's products or services, this poses a significant challenge across various industries. Understanding and predicting churn is crucial for businesses aiming to retain their customer base and maintain steady revenue streams. In competitive markets, acquiring new customers can be significantly more expensive than retaining existing ones. Thus, accurately predicting customer churn allows companies to implement targeted retention strategies, potentially saving significant resources and enhancing customer satisfaction. Machine learning has emerged as a powerful tool for predictive analytics in numerous fields, including customer churn prediction. Various algorithms, each with unique strengths and weaknesses, can be employed to predict churn

Despite the availability of various machine learning algorithms, selecting the most suitable model for customer churn prediction remains a challenge. Each algorithm offers distinct advantages and disadvantages, and their performance can vary significantly depending on the dataset and specific application. This research seeks to address which of this 3-machine learning algorithm - Stochastic Gradient Boosting (SGB), Support Vector Machine (SVM) and Random Forest provides the most accurate and reliable predictions for customer churn?

**Project Timeline**

| **Task** | **Description** | **Start Date** | **End Date** |
| --- | --- | --- | --- |
| Literature Review | Review existing literature on machine learning algorithms for churn prediction | 01-06-2024 | 14-06-2024 |
| Data Collection | Collect and preprocess the dataset from the specified source | 15-06-2024 | 20-06-2024 |
| Model Selection | Select the machine learning algorithms to be compared (Stochastic Gradient Boosting, Random Forest, SVM) | 21-06-2024 | 25-06-2024 |
| Model Implementation | Implement the selected algorithms using the dataset | 26-06-2024 | 10-07-2024 |
| Model Evaluation | Evaluate the performance of each algorithm using appropriate metrics | 11-07-2024 | 20-07-2024 |
| Result Analysis | Analyze and compare the results to determine the best-performing algorithm | 21-07-2024 | 25-07-2024 |
| Report Writing | Prepare the final report documenting the research question, methodology, results, and conclusions | 26-07-2024 | 10-08-2024 |
| Final Review and Submission | Review the report and make necessary revisions before submission | 11-08-2024 | 15-08-2024 |

**Data Management Plan**

**Dataset Link**: [<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>]

**Overview of the Dataset:** The dataset, provided by IBM, encapsulates customer demographics, usage patterns, and churn labels from a telco company in California. It comprises information on 7,043 subscribers of home phone and Internet services. With 21 columns representing various features, including demographic details like age, gender, and marital status, as well as metrics such as Churn Score, and Customer Lifetime Value (CLTV) index, the dataset offers comprehensive insights into customer behavior and churn dynamics. It serves as a valuable resource for analyzing churn patterns and developing predictive models to optimize retention strategies.

**How the Data Meets Ethical Requirements**

GDPR Compliance: The dataset is anonymized and complies with GDPR guidelines.

UH Ethical Policy: Data usage is in accordance with UH ethical standards given its anonymity.

Permission to Use Data: The dataset provided by IBM has been utilized with explicit permission, as it is freely accessible for analytical purposes, ensuring compliance with ethical standards and legal guidelines.

**Document Control**

File Naming Convention: Files will adhere to the following naming format: projectname\_studentnumber\_version\_date. For my project, it would be: CustomerChurnPrediction\_22069644\_version\_date

**Metadata**

User Documents: A ReadMe file will be included to provide an overview of the dataset and instructions for use.

**Security and Storage**

Backups: Data will be backed up weekly to my google drive as well as my GitHub.

Sharing: Data will be shared with project stakeholders via secure links on the university canvas and website

**References**

Coussement, K., Lessmann, S. and Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, pp.27–36. doi:https://doi.org/10.1016/j.dss.2016.11.007.

Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. *Artificial Intelligence Research*, 6(2), p.93. doi:https://doi.org/10.5430/air.v6n2p93.