

Neuro Nexus Innovations

Documentation On

Credit Card Fraud Detection

Submission Date:24/02/2024

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 24/02/2024 | 1.0 | Documentation | Md Rian Ur Rahaman |

Contents:

1. Introduction

2. General Description

3. Problem Statement

4. Approach

5. Handling Imbalanced Dataset Using Random and SMOTE oversampling

6. Hyperparameter Tuning

7. Tools Used

8. Conclusion

## Abstract:

This data science project focuses on developing a comprehensive solution to predict and mitigate customer churn in a subscription-based business model. The project follows a systematic end-to-end approach, encompassing data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model selection, training, and deployment. The primary objective is to empower businesses with actionable insights to retain valuable customers and optimize revenue streams.Top of Form

The project begins with data collection from diverse sources, including customer interactions, transaction history, and demographic information. Data preprocessing involves cleaning, handling missing values, and transforming raw data into a format suitable for analysis. The EDA phase employs statistical and visual techniques to uncover patterns, correlations, and potential factors influencing customer churn.

Feature engineering is a crucial step, extracting meaningful insights from the data to enhance predictive model performance. Various machine learning algorithms, such as logistic regression, decision trees, and ensemble methods, are explored and evaluated for their ability to predict customer churn accurately. Model hyperparameter tuning and cross-validation techniques are applied to optimize the model's performance.

The project emphasizes the importance of interpretability, providing stakeholders with clear insights into the factors contributing to customer churn. Model explanations and visualizations aid in understanding the decision-making process and guide strategic decision-making.

The selected model is then deployed into a production environment, allowing real-time predictions on new data. Integration with existing systems and continuous monitoring ensure the model's effectiveness over time.

The project concludes with a comprehensive evaluation of the entire process, including model performance metrics, business impact, and insights gained.

By implementing this end-to-end data science project, businesses can proactively address customer churn, reduce revenue loss, and enhance overall customer satisfaction.

Introduction:

* 1. **Why this Level Design Document?**

**The purpose of this High-Level Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.**

**The HLD will:**

**. Present all of the design aspects and define them in detail**

**. Describe the user interface being implemented**

**. List and describe the non-functional attributes like:**

**. Security**

**. Reliability**

**. Maintainability**

**. Portability**

**. Reusability**

**. Application Compatibility**

**. Resource Utilization**

**. Serviceability**

* 1. **Scope**

**The HLD documentation presents the structure of the system, such as the database architecture, application architecture(layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.**

1. **General Description:**

**2.1 Problem Statement:**

Build a machine learning model to identify fraudulent credit card Transactions. Preprocess and

normalize the transaction data, handle class imbalance issues, and split the dataset into training and

testing sets. Train a classification algorithm, such as logistic regression or random forests, to classify

transactions as fraudulent or genuine.

Evaluate the model' s performance using metrics like precision, recall, and F1-score, and consider

techniques like oversampling or under-sampling for improving results.

**2.2 Product Perspective:**

I create an end-to-end project solution and web app which able to predict the premium of the personal for health insurance.

* 1. **Approach:**

The classical machine learning tasks like data exploration, data cleaning, feature engineering, model building and model testing. Trying different machine learning algorithms that’s best fit for the above case.

Some Famous Algorithms like linear regression, Decision Tre Regression and Gradient Boosting, XG-boost Regression were used in this project.

* 1. **Create API or User Interface:**

I created a user interface using HTML, CSS and Flask.

* 1. **Handling The Imbalanced Dataset:**

**Random Oversampling:**

**Definition:**

Random oversampling is a technique used to address class imbalance in machine learning datasets by increasing the number of instances in the minority class. This is done by randomly duplicating instances from the minority class until a more balanced distribution is achieved.

**Process:**

Randomly select instances from the minority class.

Duplicate these instances to create additional copies.

Add the duplicated instances back to the training dataset.

**SMOTE Oversampling:**

**Definition:** SMOTE is a more sophisticated technique for addressing class imbalance. Instead of duplicating existing minority class instances, SMOTE generates synthetic instances to balance the class distribution. It aims to provide the model with new and informative samples for the minority class.

**Process:**

For each minority class instance, select k nearest neighbours.

Create synthetic instances along the line segments joining the instance with its neighbours.

Add these synthetic instances to the training dataset.

* 1. **Hyperparameter Tuning:**

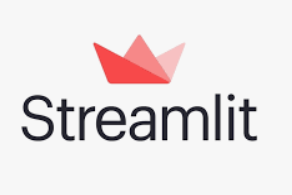
Hyperparameter tuning is a crucial step in optimizing the performance of machine learning models, including the XG-Boost classifier. XG-Boost is an efficient and scalable implementation of gradient boosting, which is an ensemble learning technique. The XG-Boost algorithm is known for its speed and performance, making it popular in various machine learning competitions and real-world applications.

Hyperparameters are configuration settings for a model that are not learned from the data but need to be specified prior to training. Tuning these hyperparameters involves finding the optimal combination of values that result in the best model performance.

**2.7 Tools Used:**

**Python** programming language and frameworks such as NumPy, Pandas, Scikit-Learn are used to build the whole model.



**Conclusion:**

This project involves building a machine learning model to detect fraudulent credit card transactions. The dataset is pre-processed, normalized, and balanced using SMOTE to handle class imbalance. A Random Forest Classifier is trained on the resampled data. Performance metrics such as precision, recall, and F1-score are calculated. The use of oversampling techniques like SMOTE helps mitigate the class imbalance issue, leading to a more robust model. The choice of Random Forest as a classifier is just one option, and other algorithms like logistic regression can also be explored based on the specific characteristics of the data. The evaluation metrics provide insights into the model's ability to correctly classify fraudulent and genuine transactions. Regular monitoring and updating of the model are recommended to adapt to changing patterns of fraud in credit card transactions.