CITATION REPORT

TOPIC: Developing a Predictive Model for GST

• Introduction

The GST Predictive Model project was undertaken as part of the "Online Challenge for Developing a Predictive Model in GST" hackathon. The primary objective of this project is to develop a machine learning model that accurately predicts whether the target parameter belongs to (class 0) or (class 1) based on the input parameters. The dataset provided consists of approximately 900,000 anonymised records, each with 21 features.

This project is crucial in helping government bodies and tax authorities streamline their analysis, ensuring a more efficient and data-driven approach to segregate the target parameter into 2 different classes based on the given input parameters.

To achieve this, we explored various machine learning techniques, with a particular focus on CatBoost and K-Nearest Neighbours (KNN) due to their effectiveness in handling imbalanced data. After rigorous preprocessing (handling missing values, scaling, and splitting the dataset), we trained the model and evaluated it on multiple metrics. The final model offers robust performance with high accuracy and precision, ensuring it can make reliable predictions on unseen data for the organising body.

• Libraries Used

1. Pandas

Purpose: Used for data manipulation and analysis also to handle the dataset, including importing, cleaning, and preprocessing.
Source: https://pandas.pydata.org/about/index.html

2. NumPy

Purpose: Fundamental package for scientific computing with Python. Used for array operations and random number generation for null value imputation.
Source: https://numpy.org/about/

3. Scikit-Learn

Purpose: Primary library for machine learning in Python. Used for KNN model, preprocessing (StandardScaler), and evaluation metrics (accuracy, precision, recall, F1 score, AUC-ROC, confusion matrix, log loss and balanced accuracy).

Source: https://scikit-learn.org/stable/about.html

4. CatBoost

Purpose: Generally used for gradient boosting on decision trees. In our case it is used for training the CatBoost classifier model.

Source: https://catboost.ai/

5. Seaborn

Purpose: Data visualisation. Used to generate visual aids such as confusion matrix plots and ROC curves.

Source: https://seaborn.pydata.org/index.html

• Imputation Techniques

1. KNN Imputation

Purpose: K-Nearest Neighbours (KNN) imputation fills missing values by averaging the four nearest rows, preserving local data patterns for more accurate predictions.

Source: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

2. Iterative Imputation

Purpose: Iterative imputation models missing values using other features, updating them iteratively to refine estimates, improving prediction accuracy for complex, correlated data.

Source: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

• Dataset

Dataset provided by the GSTN as part of the problem statement for the Hackathon.

Source: https://event.data.gov.in/challenge/online-challenge-for-developing-a-predictive-model-in-gst/

Plagiarism Declaration

We declare that the code and methodologies submitted are our original work, except where explicit citations have been provided. The project was developed in accordance with academic integrity guidelines, and we acknowledge all references to external sources in this report. Any collaborative work has been appropriately cited.

Team ID: GSTN_821
Team Members:
Rotesh Kumar
Roushan Kumar
Saumya Raj Singh
Suyash Sharthi