Citation Report For Predictive Model In GST

This machine learning model predicts target values based on a set of input features using a combination of techniques from relevant research papers and open-source libraries.

Citations:

1. Data Preprocessing: The data preprocessing steps are mainly initiated by removing the unwanted columns **Column5, Column9** and **Column14** which are not majorly correlated to any of the important features of the given dataset. We have utilised the **pandas** library for the major chunk of data processing and manipulation.

a) Navas CK - Had done analysis and identified the missing values distribution and correlation among the data. Finalised the columns to remove and to keep even finding missing values in them.
b) Sanoobar shan A - Done data preprocessing and removed unwanted columns.

2. Feature Engineering: We have analysed the columns to handle **missing values** and selected multiple methods to fill it, on which pandas dataframe functionality **Interpolate** is used to fill **column0**, **column6**, **column8** and **column15** because of its low rate of missing values on each. A Machine learning model(**LinearRegression**) is trained to **predict column3** and **column4** to fill the missing values.
The given data is found to be highly imbalanced while analysing the categorical distribution of the target column. This **data imbalance** is handled using the **SMOTE Oversampling** method.

c) Navas CK - Analysed each feature properties and distribution among the dataset. Tried different charting methods to picturise the features and given suggestions.
d) Sanoobar shan A - Tried different methods to find the missing values and implemented the final pipeline. Created a pipeline to handle data imbalance and finalised a method to go forward.

3. Model Selection: Our model selection process is mostly based on the finding from the research on Kaggle community, Informations from the AI Chat Models and also identifying the problem to be solved is classification, the data distribution and by considering the evaluation metrics of each ML model we had come to a conclusion to go with XGBoost ML algorithm to train the model.

e) Sanoobar shan A - Research and analysis on different ML models and their performances.

4. Model Training and Evaluation: We have trained the processed data with the selected ML model. By running several trial runs on the target prediction with the trained model have integrated and removed some feature engineering techniques like [Excluded]: Outlier removal, Feature scaling, .. [Included]: Interpolate, .. .

f) Sanoobar Shan A - Tried different ML models to find the perfect fit for the problem. Examined the evaluation matrices and gave feedback.

g) Navas CK - Changed or Included needed improvements or methods to influence the evaluation metrics and measured the scores. Repeated the data cycle until the best performance.

Plagiarism Declaration:

The Target Value Prediction Model is an original work created by Sanoobar Shan A and Muhammed Navas CK. All relevant research papers, libraries, and sources that have contributed to the development of this model have been properly cited and referenced. We declare that this work is original and does not infringe on any existing publications or intellectual property rights.

References:

1. https://www.kaggle.com/
2. https://www.analyticsvidhya.com/
3. Al Assistances - Llama3-70b, Claude 3.5 Sonnet.