GST Analytics Hackathon – Citation Report

Relevant Work and Libraries Cited:

1. **Scikit-learn (Pedregosa et al., 2011):**

   The scikit-learn library was extensively used for implementing machine learning models and preprocessing techniques, including the **IterativeImputer** and model evaluation metrics such as **Accuracy**, **Precision**, **Recall**, **F1 Score**, **AUC-ROC**, and **Log Loss**. This library was also utilized for building the **VotingClassifier** and **StackingClassifier** ensemble models.

   - Citation: Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 12, pp. 2825-2830, 2011.

2. **XGBoost (Chen & Guestrin, 2016):**

   XGBoost was used as one of the primary gradient-boosting models for classification. The model demonstrated impressive performance and was tuned using the **Optuna** optimization framework.

   - Citation: Chen, T. & Guestrin, C., "XGBoost: A Scalable Tree Boosting System," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794, 2016.

3. **LightGBM (Ke et al., 2017):**

   LightGBM was employed as another gradient-boosting model. It was selected for its efficiency in handling large datasets and was optimized for better model performance using **Optuna**.

   - Citation: Ke, G. et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," In Advances in Neural Information Processing Systems (NeurIPS), pp. 3146–3154, 2017.

4. **Optuna (Akiba et al., 2019):**

   Optuna was used for hyperparameter optimization in both XGBoost and LightGBM models. This framework efficiently searched for optimal hyperparameters to improve the performance of each model.

   - Citation: Akiba, T. et al., "Optuna: A Next-generation Hyperparameter Optimization Framework," In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19), pp. 2623–2631, 2019.

5. **imbalanced-learn (Lemaître, et al., 2017):**

The **imbalanced-learn** library was used for applying **SMOTE** (Synthetic Minority Over-sampling Technique) to address the class imbalance in the dataset, which generated synthetic samples for the minority class and improved model performance in classification tasks.

- Citation: Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." Journal of Machine Learning Research, 18(17), pp. 1-5.

6. **SMOTE (Chawla et al., 2002):**

   SMOTE (Synthetic Minority Over-sampling Technique) was applied to address the class imbalance issue in the dataset. This technique generated synthetic samples of the minority class, leading to improved model performance, especially in Recall and Precision.

   - Citation: Chawla, N. et al., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 16, pp. 321-357, 2002.

7. Matplotlib and Seaborn:

These libraries were used for data visualization throughout the analysis, including plotting performance metrics, confusion matrices, and calibration plots.

- Citation: Hunter, J. D., "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, 9(3), pp. 90-95, 2007.

8. Pandas (McKinney, 2010):

Pandas was utilized for data manipulation, including handling missing values, merging datasets, and preparing data for model input.

- Citation: McKinney, W., "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference (SciPy), pp. 56-61, 2010.

9. NumPy (Oliphant, 2006):

NumPy was used for handling numerical data and performing operations on arrays, essential in preparing the data for machine learning models.

- Citation: Oliphant, T., "A Guide to NumPy," USA: Trelgol Publishing, 2006.

Plagiarism Declaration:

I, **Sarang Dave** (Team ID: GSTN_539) hereby declare that the work presented in this project, including the model development, methodology, and performance analysis, is my original work. Any external sources and libraries used in this work have been appropriately cited and referenced. This submission complies with academic integrity policies, and no part of this work has been copied or plagiarized from unacknowledged sources.