In addressing the challenge of imbalanced data, a comprehensive approach was undertaken to enhance the predictive performance of the model. The first step involved mitigating class imbalance through the application of Synthetic Minority Over-sampling Technique (SMOTE), a method designed to generate synthetic instances of the minority class, thereby balancing the dataset. Subsequently, various feature selection techniques were employed to identify the most relevant attributes for model training. Techniques such as Random Forest feature importance, Lasso feature importance, permutation importance, and Gradient Boosting feature importance were applied. The selection of the optimal features was determined by calculating the average importance across these methods, ensuring a robust and representative subset for model training.

The second phase of the project focused on model selection and optimization. A diverse set of machine learning models were trained on the pre-processed data, encompassing algorithms from different families such as decision trees, support vector machines, and ensemble methods. The models were evaluated based on their F1 scores, a metric that balances precision and recall, providing a comprehensive assessment of classification performance. The top-performing model, determined by the highest F1 score, was selected for further refinement.

In the final phase, the chosen model underwent fine-tuning to optimize its hyperparameters and enhance overall performance. This iterative process involved adjusting parameters to achieve the best possible balance between precision and recall. The resulting model not only addressed the initial challenge of imbalanced data but also demonstrated superior predictive capabilities. The project's findings and methodology were consolidated into a detailed project report, outlining the steps taken from data pre-processing to model selection and fine-tuning. This report serves as a comprehensive documentation of the strategies employed and their impact on achieving a robust and effective predictive model.

**The Dataset included** **Following Columns** : ['CIF', 'CUS\_DOB', 'AGE', 'CUS\_Month\_Income', 'CUS\_Gender',

'CUS\_Marital\_Status', 'CUS\_Customer\_Since', 'YEARS\_WITH\_US',

'# total debit transactions for S1',

'# total debit transactions for S2',

'# total debit transactions for S3', 'total debit amount for S1',

'total debit amount for S2', 'total debit amount for S3',

'# total credit transactions for S1',

'# total credit transactions for S2',

'# total credit transactions for S3', 'total credit amount for S1',

'total credit amount for S2', 'total credit amount for S3',

'total debit amount', 'total debit transactions', 'total credit amount',

'total credit transactions', 'total transactions', 'CUS\_Target',

'TAR\_Desc', 'Status']

**Step 1 : EDA**

**Step2 : Data preparation and Feature Selection**

* Deleted Cols **CIF** , **CUS\_DOB , CUS\_Customer\_Since , Total \_Debit\_transasctions , Total\_Credit\_Transactions , Total\_transactions , TAR\_Desc, CUS\_Target**
* Dropped null values , only 11 null values were present in **CUS\_Month\_Income**
* The Dataset was imbalanced . Used Smote Oversampling

Tried out Various feature selection techniques and calculated the average of all the feature importance : (USED for both One hot Encoding and Ordinal Encoder and selected the best out of it)

* Random forest Feature importance
* Gradient Boosting Feature importance
* Permutation Importance
* SHAP ( shapley additive explanations)
* RFE – recursive feature elimination

**Step 3 : Modelling**

Similarly Performed the same steps as done in feature Selection , trained all the models on the data and selected the best model as final Model (Extra Tree Classifier

**Step 4 : Final Model**

Utilized Columns Transformer for feature transformation and encoding, incorporating the preprocessing steps. Employed a Pipeline for streamlined model training. Selected the Extra Tree Classifier, fine-tuned parameters, and achieved high performance metrics:

* **Training Accuracy:** 100%
* **Testing Accuracy:** 92%
* **F1 Score:** 93%
* **Precision:** 90%
* **Recall:** 95%