Credit Risk Modeling Documentation

Approach Overview

The goal of this analysis was to build a predictive model for credit risk, using a dataset with a highly imbalanced distribution of defaulters and non-defaulters. The following steps were followed:

Data Preparation

- **Data Loading:** The dataset was loaded using Pandas.
- **Data Cleaning:**
  - Columns with 100% missing values were dropped.
  - Numeric columns were imputed using the median.
- **Feature Engineering:**
  - Transaction and bureau attributes were aggregated into total, mean, and standard deviation metrics.
  - Anomalies were identified using the Isolation Forest algorithm.
- **Imbalanced Handling:**
  - Small categories (fewer than 100 records) were grouped into an 'Others' category.

Target Variable Analysis

- The target variable `bad_flag` is highly imbalanced, with 98.58% being non-defaulters (0) and only 1.42% defaulters (1).
- The distribution was visualized using a count plot.

Model Selection

Several algorithms were evaluated:

- **Logistic Regression:** Baseline linear model.
- **Random Forest:** Ensemble learning model.
- **XGBoost:** Gradient boosting method.

Random Forest performed the best among the models tested, showing high recall and precision.

Data Imbalance Management Techniques

- **Resampling:**
  - **SMOTE:** Oversampling the minority class to balance the dataset.
  - **Undersampling:** Randomly reducing the size of the majority class.
- **Class Weights:** Adjusting the class weights for models to focus on the minority class.

Feature Engineering Insights

- Categories with fewer records were grouped under 'Others.'
- Bad accounts were more concentrated in lower transaction values.
- Bad accounts exhibited a higher number of bureau inquiries.
- Attributes such as `onus_attribute_1` showed distinct differences between bad and good accounts.

Algorithm Descriptions

- **Logistic Regression:** A linear model suitable for binary classification.
- **Random Forest:** An ensemble of decision trees that reduces overfitting by averaging multiple models.
- **XGBoost:** Gradient boosting technique that improves predictive performance by sequentially training models.

Performance Metrics

- **Confusion Matrix:** Analyzed True Positive, True Negative, False Positive, and False Negative rates.
- **Precision, Recall, F1-Score:** Focused on minority class detection.
- **ROC-AUC Score:** Assessed overall model performance.

Model Evaluation Results

Logistic Regression:

- Accuracy: 71%
- ROC-AUC: 0.77

Random Forest:

- Accuracy: 99%
- ROC-AUC: 0.999

XGBoost:

- Accuracy: 99%
- ROC-AUC: 0.998

Key Insight: Random Forest outperformed other models and was chosen for further predictions due to its high recall and accuracy.

Key Insights and Observations

Insights:

- **Target Variable Imbalance:**

  The dataset exhibited a highly imbalanced distribution of the target variable (bad_flag), with 98.58% non-defaulters and only 1.42% defaulters. This highlighted the need for techniques to handle data imbalance effectively.

- **Feature Analysis:**
  - Bad accounts were predominantly associated with lower transaction values and exhibited higher bureau inquiries.
  - Certain attributes, such as onus_attribute_1, displayed significant differences between defaulters and non-defaulters.
- **Feature Engineering:**
  - Aggregation of transaction and bureau attributes into metrics such as total, mean, and standard deviation provided deeper insights into account behaviors.
  - Identification of anomalies using the Isolation Forest algorithm improved data quality and feature refinement.
- **Model Performance:**
  - Among the models tested, Random Forest achieved the highest performance with an accuracy of 99% and a ROC-AUC score of 0.999, surpassing Logistic Regression and XGBoost.
  - Handling of imbalanced data through SMOTE and undersampling played a crucial role in improving minority class detection.
- **Correlations:**

  Strong correlations among certain features suggested redundancy, requiring careful feature selection to improve model performance.

Inferences:

- **Credit Risk Identification:**

  Accounts with lower transaction values and higher bureau inquiries are more likely to default, offering a clear risk profile for early detection.

- **Model Effectiveness:**

  Random Forest's robustness against overfitting and ability to manage complex interactions between features makes it an ideal choice for this imbalanced dataset.

- **Business Application:**

  The insights derived from this analysis enable financial institutions to identify high-risk accounts early, thereby implementing effective risk mitigation strategies.

- **Scalability:**

  The combination of advanced resampling techniques and feature engineering demonstrates a scalable framework for future datasets with similar imbalances.

Conclusion

Random Forest was selected as the final model due to its superior performance. The insights generated from the exploratory data analysis and feature engineering provided a comprehensive understanding of the risk patterns in the data, allowing for effective model building and evaluation.