Convolve 3.0

Team SciPy

(Abhay Gupta, Vedant Chetnabhai Joshi, Ajay Kumar Lolla)

Project Documentation

Problem Statement

The bank aims to develop a predictive model to evaluate the likelihood of default among its existing credit card customers. This model will assist in identifying high-risk individuals, enabling the bank to take proactive measures to mitigate financial losses.

Challenges

1. **High Dimensionality:**
    - The dataset contains over 1,200 features, presenting significant challenges in processing, computation, and feature selection. High dimensionality increases model complexity and the risk of overfitting.
2. **Missing Values:**
    - The presence of NULL values across multiple features poses a risk of skewing results if not addressed with appropriate imputation techniques.
3. **Class Imbalance:**
    - The target variable is heavily imbalanced, with a disproportionately smaller number of defaulters compared to non-defaulters. This imbalance can lead to biased model predictions and poor generalizability.

Pipeline and Methodology

1. **Data Analysis and Visualization:**
    - Conducted exploratory data analysis (EDA) to understand the distribution, correlations, and overall structure of the dataset. EDA also helped identify the presence of missing values and outliers.
2. **Feature Segmentation:**
    - Segregated the dataset into four logical segments based on feature relevance to enhance interpretability and processing:
        - **Onus Attributes:** Features directly related to the individual's financial responsibility.
        - **Transaction Attributes:** Features capturing spending patterns and transactional behavior.

        - **Bureau Enquiry:** Data from credit bureau inquiries.
        - **Bureau:** Comprehensive historical credit data.
3. **Handling Missing Values:**
    - Applied imputation strategies tailored to each segment to fill in missing values. For numerical features, used the mean imputation strategy (`SimpleImputer(strategy='mean')`), ensuring consistency and minimizing data loss.
    - Converted imputed arrays back into DataFrames to maintain compatibility with subsequent processing steps.
4. **Standardization:**
    - Standardized all features using `StandardScaler` to normalize the data and bring all variables to a comparable scale. This step is critical for ensuring that no single feature dominates the model due to scale differences.
5. **Dimensionality Reduction:**
    - Applied Principal Component Analysis (PCA) to each data segment, reducing dimensionality while retaining the most significant variance:
        - Selected 5 principal components per segment, resulting in a total of 20 features across all segments.
        - PCA was chosen to address the curse of dimensionality and mitigate overfitting risks, ensuring efficient computation without substantial information loss.
    - Reconstructed reduced arrays into DataFrames for ease of integration.
6. **Data Integration and Model Training:**
    - Merged the reduced datasets from all four segments into a single consolidated dataset for training.
    - Utilized a Voting Classifier to combine the predictive power of multiple algorithms, enhancing overall model robustness and performance. Voting Classifier combines base models to improve accuracy and generalization.
7. **Performance Evaluation:**
    - Evaluated the model's predictive power using:
        - **Accuracy:** Assessed the proportion of correctly classified instances.
        - **AUC-ROC Curve:** Measured the model's ability to distinguish between defaulters and non-defaulters, emphasizing its effectiveness in handling class imbalance.

Results

- **AUC (Area Under the Curve):** 0.77, indicating good discriminatory power.
- **Accuracy:** 92%, demonstrating the model's reliability in correctly classifying credit card customers as defaulters or non-defaulters.

Conclusion

The proposed pipeline successfully addresses the challenges of high dimensionality, missing values, and class imbalance. By leveraging PCA for dimensionality reduction, appropriate imputation strategies, and an ensemble-based Voting Classifier, the model achieves significant predictive performance. The insights from this model can empower the bank to implement effective risk mitigation strategies and optimize its credit portfolio management.