**Handling Imbalanced Data in Classification Using Logistic Regression**

**Mesut Erkin Ozokutgen**  
**Student ID: 20230945**

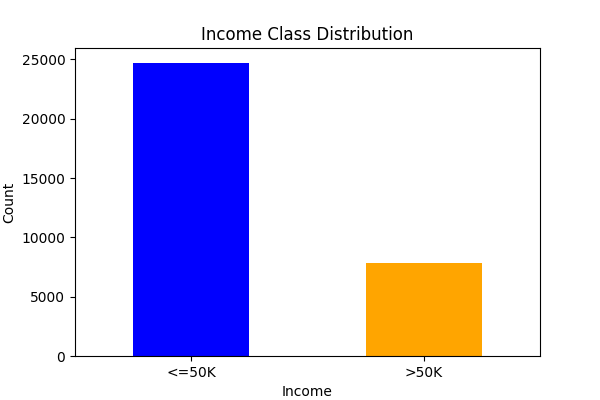
**1. Introduction**

This report presents an experimental analysis on handling imbalanced datasets in classification using the Logistic Regression model. The objective is to evaluate the performance of the model using different resampling strategies on the **UCI Adult dataset**, which is widely used for income classification. We assess performance using Accuracy, Recall, Precision, Specificity, ROC-AUC, and F1-score. Additionally, we explore dimensionality reduction via Linear Discriminant Analysis (LDA).

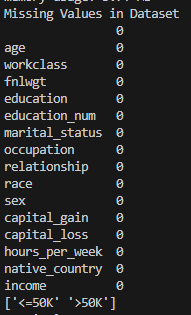
**2. Dataset Overview**

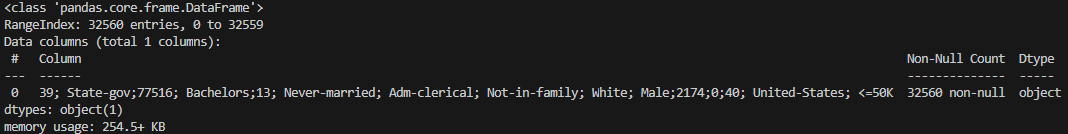
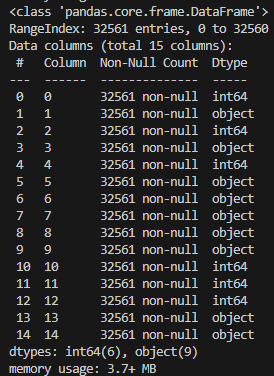
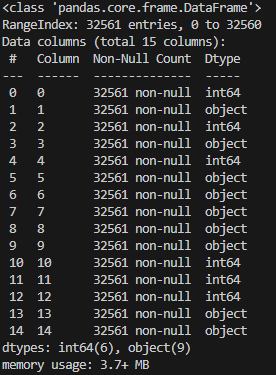
The Adult dataset contains 32,561 instances with 15 features, both numerical and categorical. The target variable is **income**, categorized as <=50K and >50K. The class distribution is imbalanced, with significantly more <=50K instances.

* No missing values were found in the dataset.
* After initial cleaning and formatting, categorical features were one-hot encoded, and numerical features were normalized to the interval [-1, 1].

**Class Distribution Visualization**: ****

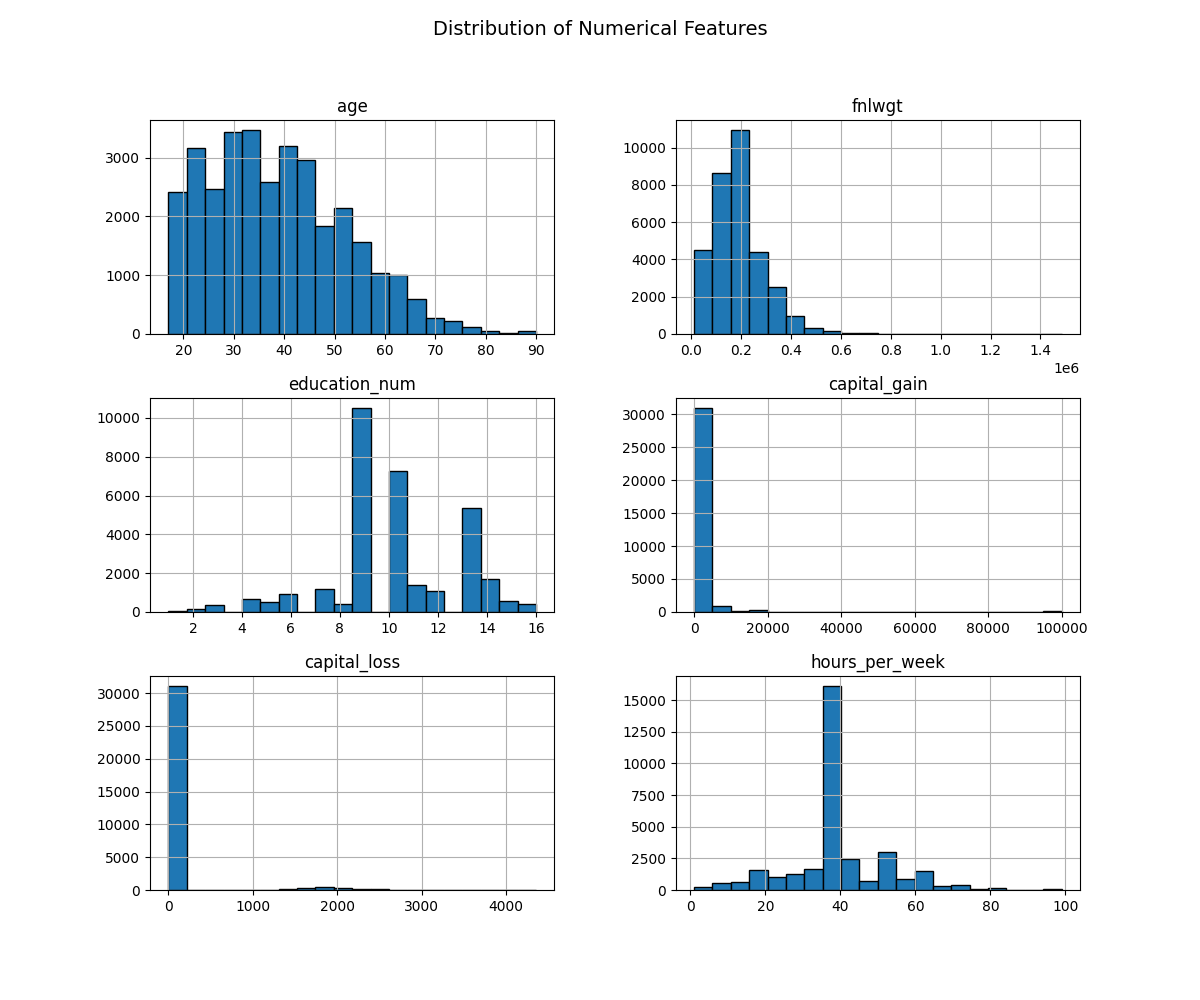
**Missing Value Check**:



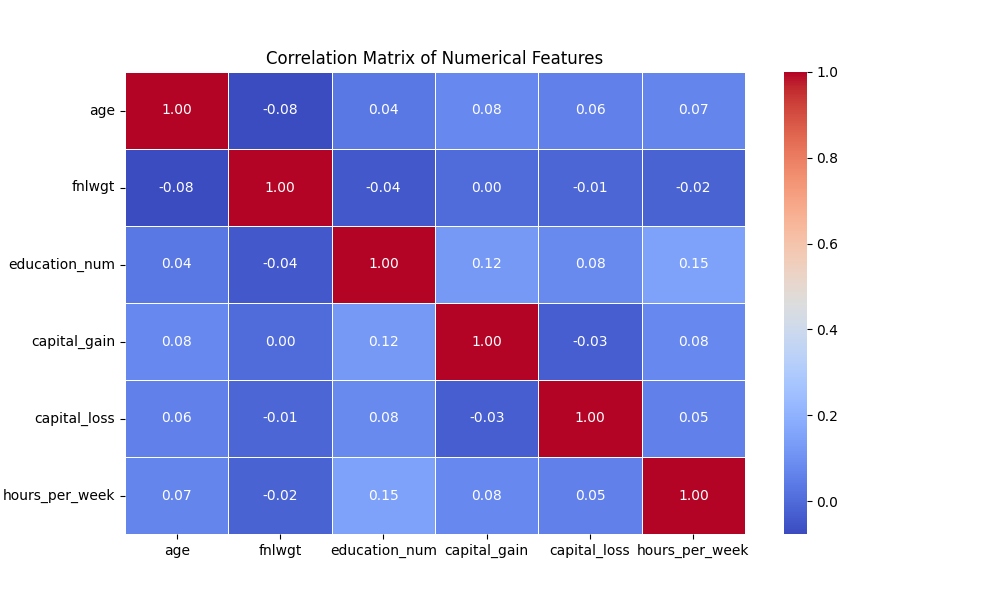
**Raw Data Format and Data Types**:  

**3. Exploratory Data Analysis**

A univariate distribution of all numerical features was plotted. The data shows right-skewed patterns in capital\_gain and capital\_loss, and a peak around 40 hours for hours\_per\_week.

**Feature Distribution Histograms**: 

A correlation heatmap indicates weak correlations among features. The highest observed correlation is between education\_num and hours\_per\_week (0.15).

**Correlation Matrix**: 

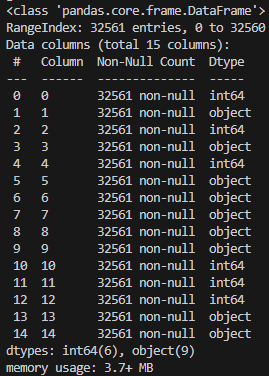
**4. Baseline Logistic Regression Model**

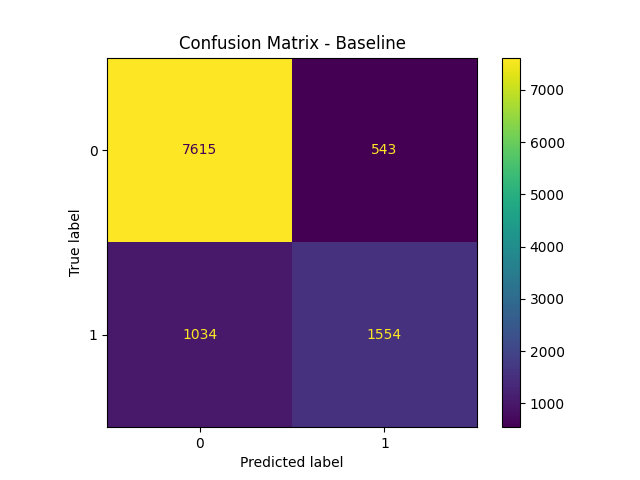
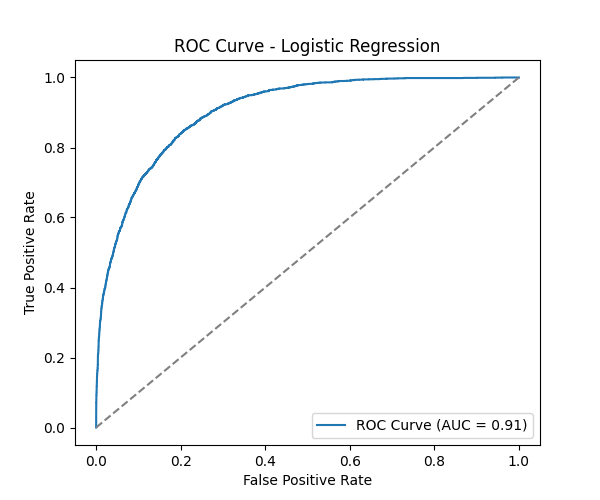
The Logistic Regression model was first trained without resampling. It was evaluated on a hold-out test set.

**Performance Metrics**:

* Accuracy: 0.853
* Recall: 0.600
* Precision: 0.741
* Specificity: 0.933
* F1-score: 0.663
* ROC-AUC: 0.906

**Metrics Table**:



**Confusion Matrix (Baseline)**:   
**ROC Curve (Baseline)**: 

**5. Resampling Techniques and Impact**

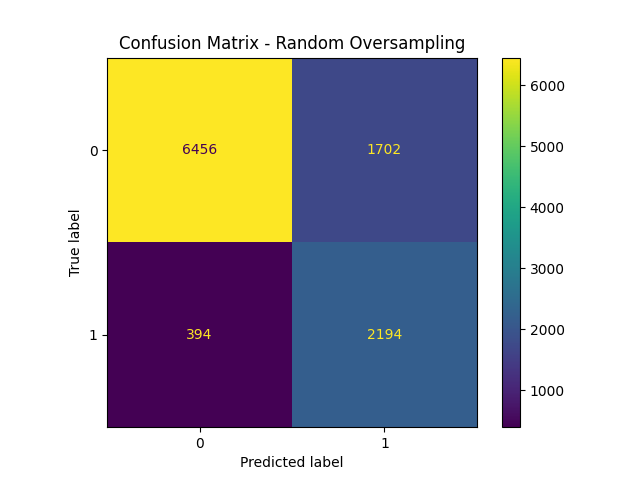
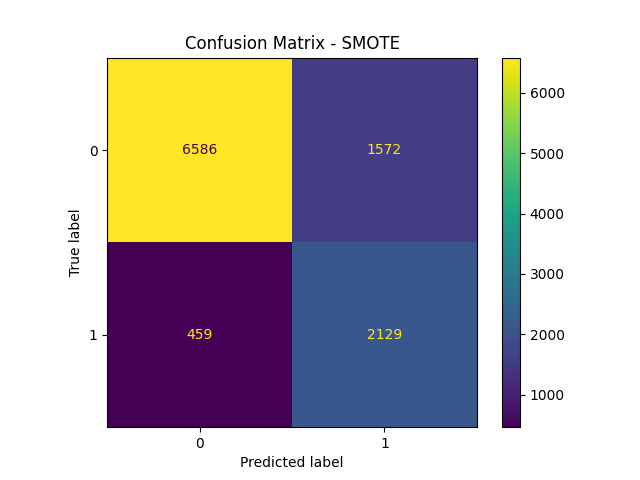
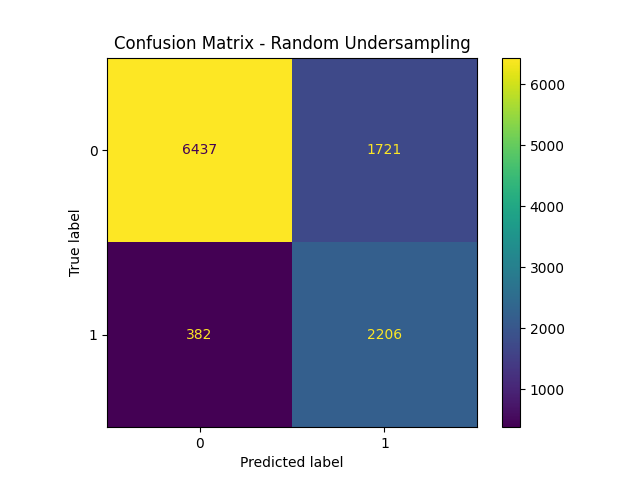
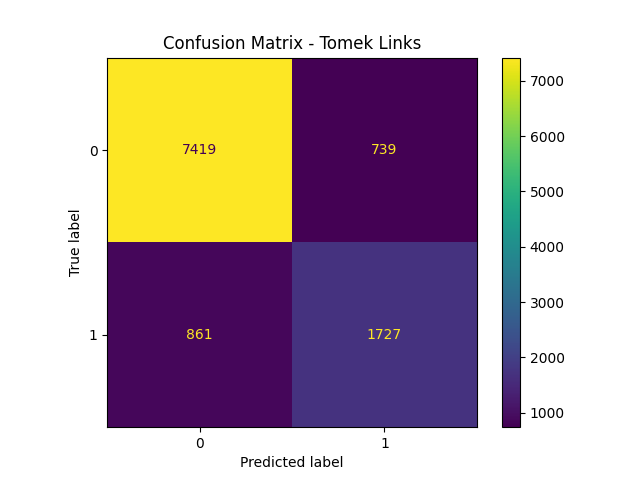
To address class imbalance, we applied the following resampling methods:

* **Random Oversampling**
* **SMOTE** (Synthetic Minority Oversampling Technique)
* **Random Undersampling**
* **Tomek Links**

Each method was evaluated with Logistic Regression.

**Performance Summary Table**: [Adsız6.png]

**Confusion Matrices**:

* Random Oversampling: 
* SMOTE: 
* Random Undersampling: 
* Tomek Links: 

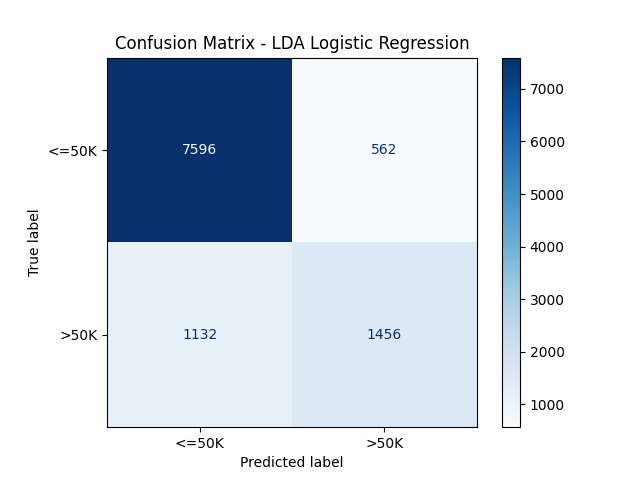
From the results, Random Undersampling and SMOTE yielded high recall (~85%) while maintaining a balanced F1-score (~0.67). However, precision slightly dropped in oversampling methods.

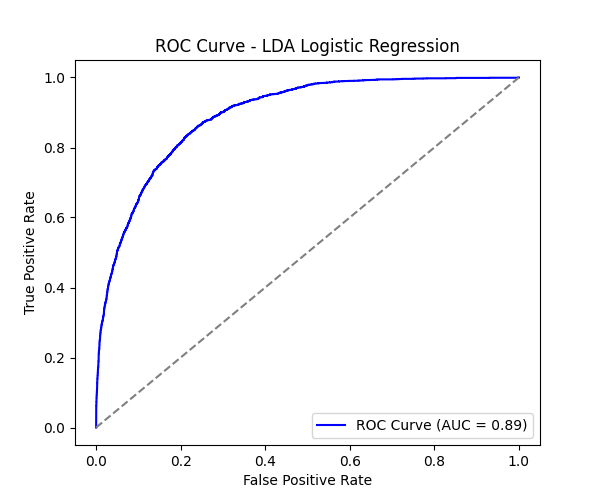
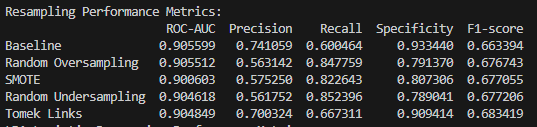
**6. Dimensionality Reduction Using LDA**

We applied **Linear Discriminant Analysis (LDA)** to reduce features to a single discriminative component, and retrained Logistic Regression.

**LDA Model Performance**:

* Accuracy: 0.842
* Recall: 0.563
* Precision: 0.722
* Specificity: 0.931
* F1-score: 0.632
* ROC-AUC: 0.894

**LDA Confusion Matrix**: 

**ROC Curve (LDA)**:   
**Metrics Table (LDA)**: 

**7. Conclusion**

* The dataset exhibits moderate class imbalance (~76% to 24%).
* Logistic Regression performs well with ROC-AUC ~0.91.
* SMOTE and undersampling methods improved recall significantly, making them suitable for imbalanced classification.
* LDA effectively reduced dimensionality with minimal loss in classification performance.

These results suggest that simple resampling techniques paired with Logistic Regression can deliver robust results even in imbalanced settings.