**Logistic Regression Case Study Report**

**Project Title: Income Classification Using Census Data**

**Objective**

The goal of this project is to build a **binary classification model** using **logistic regression** to predict whether an individual earns **more than $50K per year** based on their demographic and employment information from U.S. Census Bureau data.

**Dataset Description**

* **Source:** 1994 U.S. Census Bureau data
* **Records:** ~32,000
* **Target:** annual\_income — whether an individual earns >50K (1) or <=50K (0)
* **Features:** 15+ including age, education, occupation, hours per week, etc.

**Exploratory Data Analysis (EDA)**

**1. Target Variable Distribution**

| **Income Category** | **Count** | **Percentage** |
| --- | --- | --- |
| ≤50K | ~24,000 | ~75% |
| >50K | ~8,000 | ~25% |

**Observation**: The dataset is **imbalanced**, with the majority of individuals earning ≤50K.

**2. Missing Values & Cleaning**

* Missing values encoded as ? in workclass, occupation, native-country.
* Replaced with "others" for consistency and to retain all records.

**3. Feature Significance Tests**

**a. Categorical Variables – Chi-Square Test**

| **Feature** | **Associated with Income?** |
| --- | --- |
| workclass | ✅ Yes |
| education | ✅ Yes |
| marital-status | ✅ Yes |
| occupation | ✅ Yes |
| relationship | ✅ Yes |
| race | ❌ No |
| sex | ✅ Yes |
| native-country | ✅ Yes |

**Only features significantly associated with income were retained**.

**b. Continuous Variables – Two-Sample Z-Test**

| **Feature** | **p-value** | **Retain?** |
| --- | --- | --- |
| age | < 0.001 | ✅ Yes |
| education-num | < 0.001 | ✅ Yes |
| capital-gain | < 0.001 | ✅ Yes |
| capital-loss | < 0.001 | ✅ Yes |
| hours-per-week | < 0.001 | ✅ Yes |

These continuous features **differ significantly** between high and low-income groups.

**Data Preprocessing**

* **Dropped** irrelevant column fnlwgt.
* **Encoded categorical variables** using **one-hot encoding** (with drop\_first=True to avoid multicollinearity).
* **Final shape:** All features are numeric, and dataset is ready for modeling.

**Feature Multicollinearity Check (VIF)**

| **Feature** | **VIF** |
| --- | --- |
| age | ~1.2 |
| education-num | ~1.6 |
| capital-gain | ~1.3 |
| capital-loss | ~1.1 |
| hours-per-week | ~1.4 |

**No multicollinearity issue** (VIF < 5 for all continuous predictors)

**Model Development: Logistic Regression**

**Model Specification**

* **Algorithm:** Binary Logistic Regression
* **Train-Test Split:** 80:20
* **Library Used:** statsmodels.api.Logit

mod = sm.Logit(y\_train, x\_train).fit()

y\_pred = mod.predict(x\_test)

y\_pred\_class = (y\_pred >= 0.5).astype(int)

**Classification Report**

| **Metric** | **Class 0 (≤50K)** | **Class 1 (>50K)** |
| --- | --- | --- |
| Precision | 0.88 | 0.73 |
| Recall | 0.93 | 0.60 |
| F1-Score | 0.90 | 0.66 |
| **Accuracy** | **85%** |  |

**Interpretation**:

* **High precision and recall** for class 0 (low-income individuals).
* Moderate precision for class 1 (high-income), but **recall is lower (60%)**.
* **Model misses 40% of high-income individuals**, possibly due to class imbalance.

**Key Business Insights**

**Influencing Features**

| **Feature** | **Impact on Income** |
| --- | --- |
| Education Level | Higher → More Income |
| Age | Older → Higher Likelihood of >50K |
| Hours Per Week | More Hours → More Income |
| Capital Gain | Positive Correlation |
| Occupation Type | Strong Association |
| Marital Status | Married Individuals Tend to Earn More |

**Category Distributions**

**Workclass:**

* Most individuals belong to Private sector.
* Self-employed individuals show higher income probability.

**Education:**

* Advanced education (Bachelors or higher) strongly correlates with income >50K.

**Native Country:**

* U.S.-born individuals dominate the dataset; income patterns vary across nationalities.

**Limitations**

* **Imbalanced dataset**: Only 25% individuals are >50K earners.
* **Lower recall** for >50K class may hinder real-world applications where identifying high-income individuals is crucial.
* No hyperparameter tuning or regularization used (e.g., Lasso or Ridge).
* Assumes linear relationship between log-odds and input features.

**Conclusion**

| **Model** | **Type** | **Accuracy** | **Strength** | **Weakness** |
| --- | --- | --- | --- | --- |
| Logistic Regression | Classification | 85% | Interpretable, efficient for binary tasks | Lower recall for high-income class (1) |

The model is well-suited for understanding income patterns and predicting lower-income individuals with high accuracy.