Live Tap Consultancies

@livetapcons.com

INCOME PREDICTION STUDY

Live Tap Consultancies

@livetapcons.com

TITLE: PREDICTING INCOME: A MACHINE LEARNING APPROACH

SUBTITLE: UNCOVERING KEY DRIVERS AND DISPARITIES IN INCOME PREDICTION

Introduction & Objectives

Introduction - Understanding Income

- Introduction: The Importance of Income Prediction
- Income is a fundamental economic indicator.
- Predicting income can inform policy, marketing, and resource allocation.
- Machine learning offers powerful tools to analyse complex relationships in demographic and work-related data.

Study Objectives

- **OBJECTIVE 1: PREDICTIVE MODELING:** DEVELOP MODELS TO CLASSIFY ANNUAL INCOME INTO >50K OR <=50K CATEGORIES USING DEMOGRAPHIC AND WORK-RELATED FEATURES.
- **OBJECTIVE 2: FEATURE IMPORTANCE:** IDENTIFY THE MOST INFLUENTIAL ATTRIBUTES (E.G., EDUCATION, OCCUPATION, HOURS WORKED) THAT DRIVE THE >50K INCOME OUTCOME.
- **OBJECTIVE 3: SUBGROUP ANALYSIS:** EXAMINE HOW INCOME VARIES ACROSS DIFFERENT DEMOGRAPHIC SUBGROUPS (E.G., GENDER, RACE, MARITAL STATUS, NATIVE COUNTRY).

Data Overview

Target Variable

Income (<=50K vs. >50K)

Key Features

age, education_num, capital_gain, capital_loss, hours_per_week, marital_status, occupation, relationship, gender, native_country

Class Imbalance :
We have significant imbalance present, with individuals with >50K being the minority class.

DATA PREPROCESSING

- **CATEGORICAL ENCODING:** ONE-HOT ENCODING APPLIED TO CONVERT CATEGORICAL FEATURES INTO NUMERICAL FORMAT.
- **FEATURE SCALING:** NUMERICAL FEATURES STANDARDIZED USING STANDARDSCALER TO ENSURE FAIR WEIGHTING IN LINEAR MODELS.
  - Formula: Xscaled=(X−μ)/σ.
- **TRAIN-TEST SPLIT:** DATA DIVIDED INTO TRAINING AND TESTING SETS, ENSURING STRATIFICATION OF THE TARGET VARIABLE.

MODELS USED :

LOGISTIC REGRESSION

DECISION TREE

RANDOM FOREST

FINDINGS

Logistic Regression

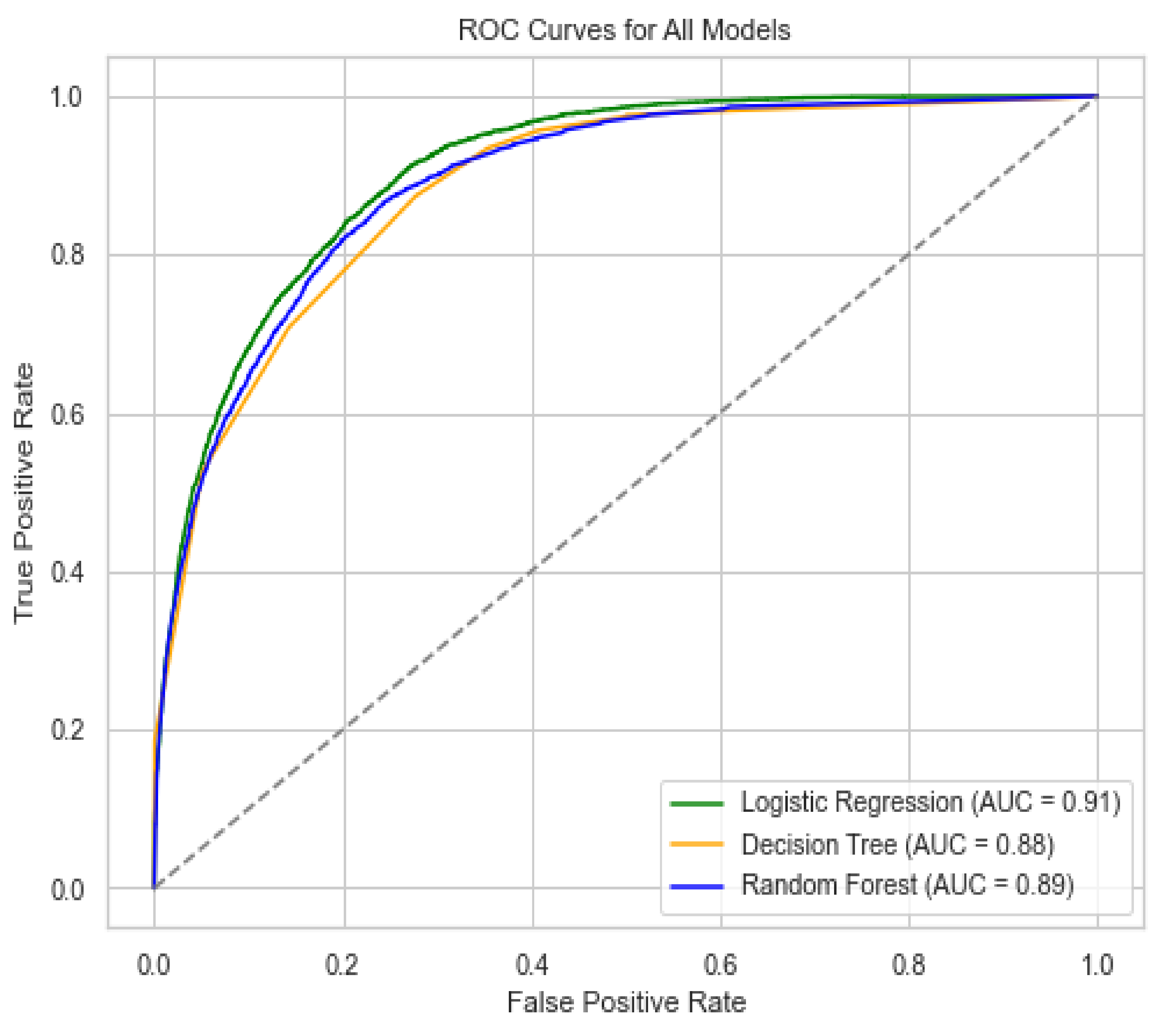
QUANTIFES LINEAR EFFECT(E.G CAPITAL_GAIN HAS THE HIGHEST POSITIVE COEFFICIENT.

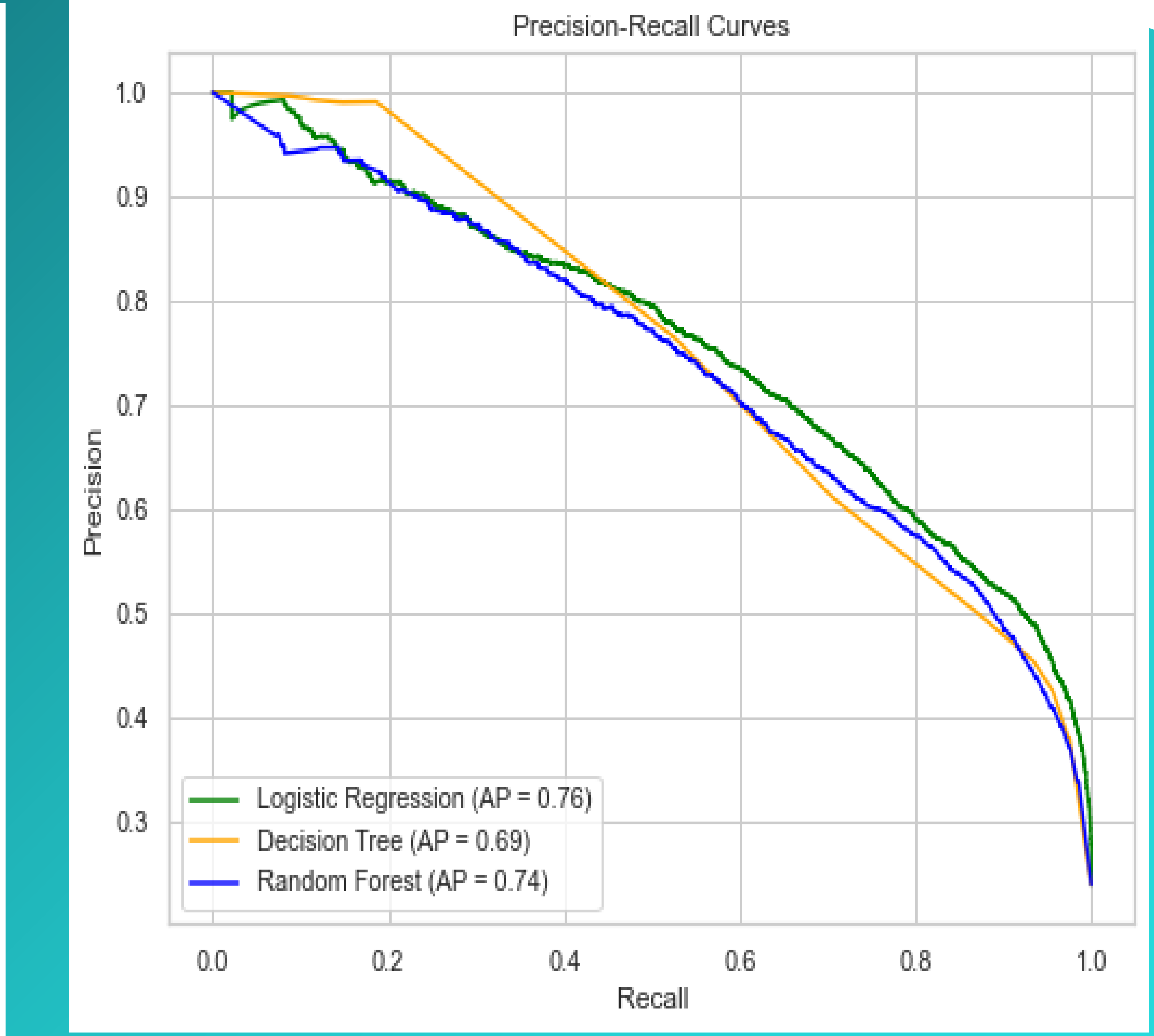
Decision Tree

EMPHASIZES SINGLE DOMINANT SPLITS (MARITAL STATUS_MARRIED-CIV-SPOUSE AS IMPORTANT THAN OTHERS)

RANDOM FOREST

PROVIDES MORE ROBUST FEATURE HIERARCHY HIGHLIGHTING AGE AND HOURS_PER _WEEK AS OVERALL MOST IMPORTANT.





IN BOTH GRAPHS LOGISTIC MODEL STANDS OUT AS THE BEST MODEL TO HELP US IN PREDICTING HIGH INCOME.
WHEREAS DECISION TREE SEE TO BE LEAST PERFORMING MODEL

RECOMMENDATIONS

1. **Prioritize Random Forest**: It shows the highest overall accuracy (0.84) and a strong balance of precision and recall for the minority class (F1-score 0.66, AP 0.74), making it the most robust model for this task.
2. **Enhance Imbalance Handling:** Explore 'class_weight' parameters or advanced 'imblearn' techniques (e.g., ADASYN) within ensemble models to further optimize minority class prediction.
3. **Conduct Deeper Subgroup Analysis:** Perform dedicated statistical analyses and use model-agnostic interpretability tools like SHAP values to fully investigate income disparities related to gender, race, and native country, as these features showed low direct importance in current models.

CONCLUSION

Our study effectively predicts income using demographic and work-related features. The **Random Forest model is the top performer**, achieving the best overall accuracy and robustly balancing precision and recall for both income categories. Key drivers of higher income are consistently identified as capital_gain, marital_status_Married-civ-spouse, education_num, age, and hours_per_week. While income varies significantly by marital status, education, and occupation, deeper analysis is needed for gender and race to fully understand disparities.