PROSPERNET

LEVERAGING MACHINE LEARNING FOR NON-PROFIT FUNDRAISING: PREDICTING HIGH-INCOME INDIVIDUALS

Janani Karthikeyan
002830003

PROJECT FLOW

01

Introduction to ProsperNet

1. Project Overview
2. Methodology

02

Data Preprocessing and Exploratory Data Analysis

1. Data Cleaning and Transformation
2. Exploratory Data Analysis

03

Model Implementation and Evaluation

1. Baseline Model Evaluation
2. Hyperparameter Tuning

04

Comparative Analysis of Machine Learning Algorithms

1. Algorithm Performance Metrics
2. Algorithm Recommendations

01 : Introduction to ProsperNet

Project Overview

The "ProsperNet: Predicting High-Income Individuals" project leverages machine learning to enhance non-profit fundraising efforts. It develops a predictive model using supervised learning algorithms, utilizing demographic and financial data from the 1994 U.S. Census to identify individuals with annual incomes exceeding $50,000. This model aids non-profits in refining donation solicitation and outreach efforts, providing a strategic advantage in fundraising activities. The data, sourced from the UCI Machine Learning Repository, undergoes meticulous preprocessing and analysis to construct an accurate and applicable model. ProsperNet empowers non-profits with data-driven insights for informed donor engagement, contributing significantly to the optimization of fundraising strategies.

01 : Introduction to ProsperNet

Methodology

- **Data Acquisition and Pre-processing:** Perform data cleaning, normalization, and splitting into training and testing sets.
- **Exploratory Data Analysis (EDA):** Initial analysis to identify patterns, correlations, and key features within the dataset.
- **Feature Selection:** Determining the most relevant features for accurate income prediction.
- **Model Fitting:** Implemented four supervised learning algorithms (Logistic Regression, AdaBoost, Decision Trees, and MLP Classifier) and evaluated baseline models using initial features without hyperparameter tuning.
- **Hyperparameter Tuning:** Optimizing and re-evaluating models to achieve peak performance.
- **Model Evaluation:** Testing the supervised learning algorithms for the best predictive accuracy and computational efficiency.

02 : Data Preprocessing and Exploratory Data Analysis

Dataset Description

The data underpinning this project is extracted from the UCI Machine Learning Repository, courtesy of contributions by Ron Kohavi and Barry Becker. This dataset, comprising a wide array of demographic and economic indicators, serves as the foundation for model training and validation.

- Number of Instances: Over 32,000 records in the training set.
- Number of Attributes: 14 features plus a target label.

Features: age, workclass, fnlwgt, education, education-level, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income (target variable).

02 : Data Preprocessing and Exploratory Data Analysis

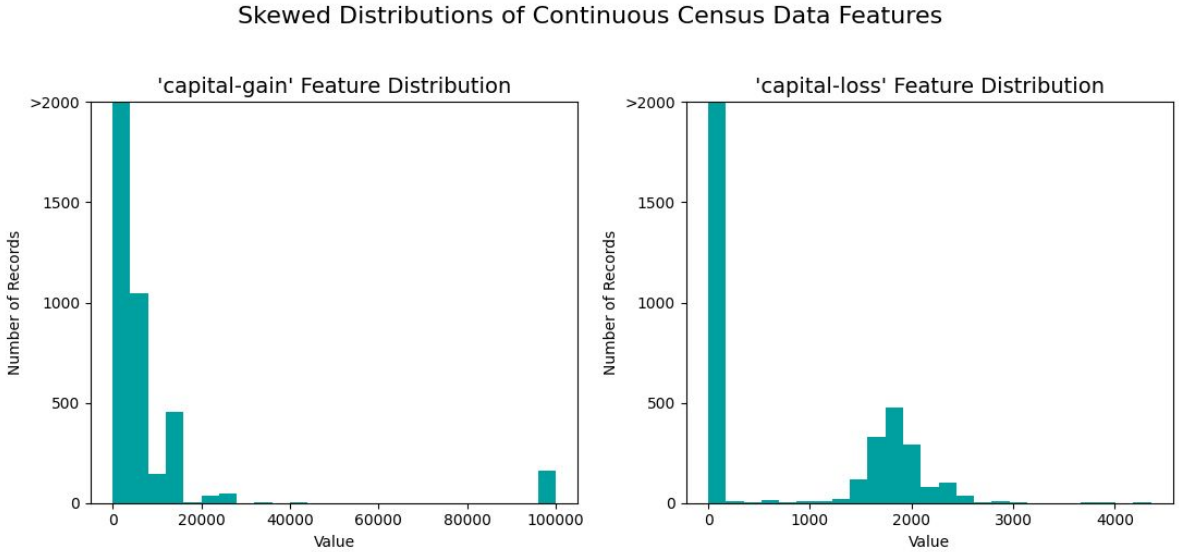
Data Cleaning & Transformation

The following steps were performed as part of the data preprocessing.

- **Data Importing & Cleaning** - using data.dropna()
- **Transforming Skewed Continuous Features** - to re-scale the high-magnitude values and reduce the effect of outliers
- **Log-transform the skewed features** - to reduce skewness in the distribution
- **Data Normalization** - MinMaxScaler is a feature scaling technique that shrinks the range of data to [0,1]
- **One-hot Encoding** - to encome the target variable “income”
- **Splitting data into training and testing sets** - training set contains 26,048 samples, and the testing set has 6,513 samples

02 : Data Preprocessing and Exploratory Data Analysis

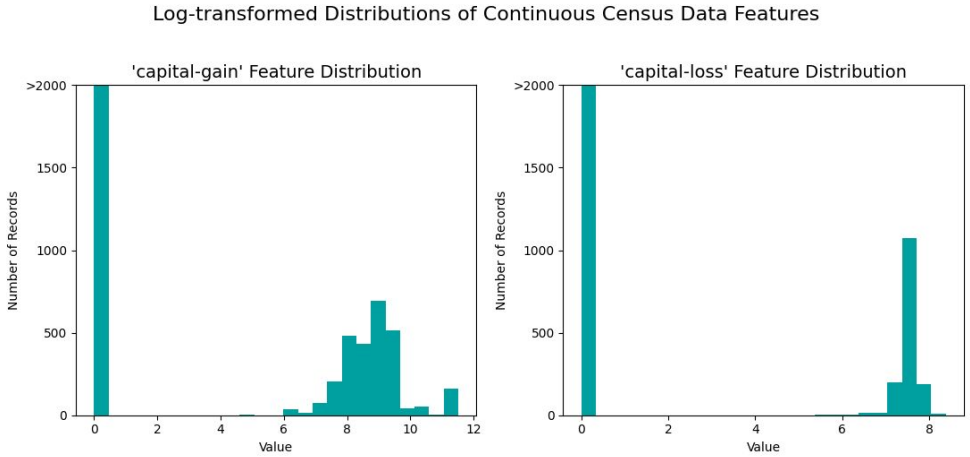
Data Transformation - Skewed Continuous Features



- The 'capital-gain' and 'capital-loss' histograms post-transformation should ideally show a more uniform or normal distribution compared to pre-transformation.
- However, from the output, there are still a large number of zeros that don’t change much with transformations like log, as log(0) is undefined.

02 : Data Preprocessing and Exploratory Data Analysis

Data Transformation - Log-transform the skewed features



- The histograms display the 'capital-gain' and 'capital-loss' data after a log transformation, which reduces their skewness.
- Values of zero are transformed to zero, resulting in a high bar at that point.
- The transformation compresses the range of non-zero values, moderating the impact of outliers and making the distribution less skewed.

02 : Data Preprocessing and Exploratory Data Analysis

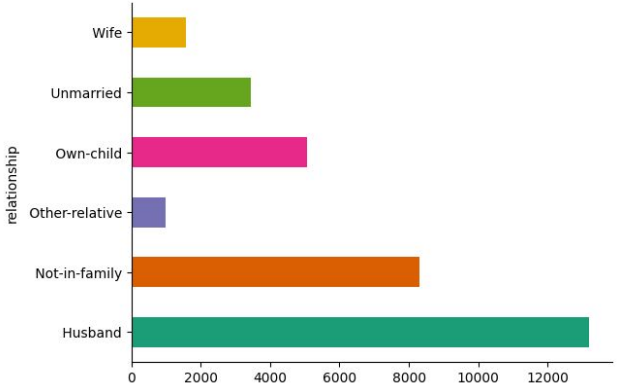
Data Transformation - One-hot Encoding

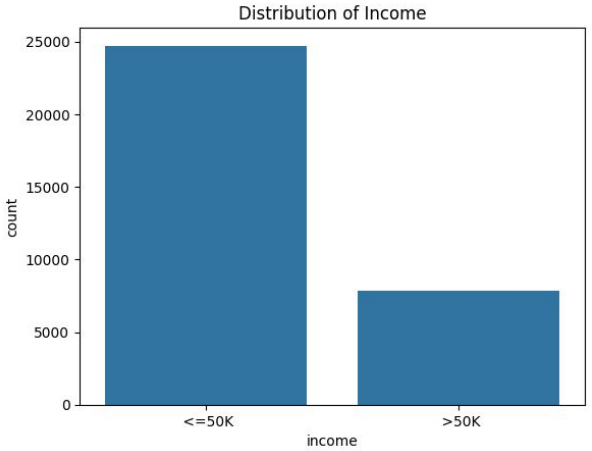
108 total features after one-hot encoding:

['age', 'fnlwgt', 'education-level', 'capital-gain', 'capital-loss', 'hours-per-week', 'workclass_ ?', 'workclass_ Federal-gov', 'workclass_ Local-gov', 'workclass_ Never-worked', 'workclass_ Private', 'workclass_ Self-emp-inc', 'workclass_ Self-emp-not-inc', 'workclass_ State-gov', 'workclass_ Without-pay', 'education_ 10th', 'education_ 11th', 'education_ 12th', 'education_ 1st-4th', 'education_ 5th-6th', 'education_ 7th-8th', 'education_ 9th', 'education_ Assoc-acdm', 'education_ Assoc-voc', 'education_ Bachelors', 'education_ Doctorate', 'education_ HS-grad', 'education_ Masters', 'education_ Preschool', 'education_ Prof-school', 'education_ Some-college', 'marital-status_ Divorced', 'marital-status_ Married-AF-spouse', 'marital-status_ Married-civ-spouse', 'marital-status_ Married-spouse-absent', 'marital-status_ Never-married', 'marital-status_ Separated', 'marital-status_ Widowed', 'occupation_ ?', 'occupation_ Adm-clerical', 'occupation_ Armed-Forces', 'occupation_ Craft-repair', 'occupation_ Exec-managerial', 'occupation_ Farming-fishing', 'occupation_ Handlers-cleaners', 'occupation_ Machine-op-inspct', 'occupation_ Other-service', 'occupation_ Priv-house-serv', 'occupation_ Prof-specialty', 'occupation_ Protective-serv', 'occupation_ Sales', 'occupation_ Tech-support', 'occupation_ Transport-moving', 'relationship_ Husband', 'relationship_ Not-in-family', 'relationship_ Other-relative', 'relationship_ Own-child', 'relationship_ Unmarried', 'relationship_ Wife', 'race_ Amer-Indian-Eskimo', 'race_ Asian-Pac-Islander', 'race_ Black', 'race_ Other', 'race_ White', 'sex_ Female', 'sex_ Male', 'native-country_ ?', 'native-country_ Cambodia', 'native-country_ Canada', 'native-country_ China', 'native-country_ Columbia', 'native-country_ Cuba', 'native-country_ Dominican-Republic', 'native-country_ Ecuador', 'native-country_ El-Salvador', 'native-country_ England', 'native-country_ France', 'native-country_ Germany', 'native-country_ Greece', 'native-country_ Guatemala', 'native-country_ Haiti', 'native-country_ Holand-Netherlands', 'native-country_ Honduras', 'native-country_ Hong', 'native-country_ Hungary', 'native-country_ India', 'native-country_ Iran', 'native-country_ Ireland', 'native-country_ Italy', 'native-country_ Jamaica', 'native-country_ Japan', 'native-country_ Laos', 'native-country_ Mexico', 'native-country_ Nicaragua', 'native-country_ Outlying-US(Guam-USVI-etc)', 'native-country_ Peru', 'native-country_ Philippines', 'native-country_ Poland', 'native-country_ Portugal', 'native-country_ Puerto-Rico', 'native-country_ Scotland', 'native-country_ South', 'native-country_ Taiwan', 'native-country_ Thailand', 'native-country_ Trinadad&Tobago', 'native-country_ United-States', 'native-country_ Vietnam', 'native-country_ Yugoslavia']

02 : Data Preprocessing and Exploratory Data Analysis

Exploratory Data Analysis





- The horizontal bar chart categorizes individuals by family relationship, showing 'Husband' as the most common, followed by 'Not-in-family', with 'Wife' as the least common among the displayed categories.
- The bar chart illustrates the income distribution, showing a larger count of individuals earning '<=50K' compared to those earning '>50K', indicating an income disparity within the dataset.

02 : Data Preprocessing and Exploratory Data Analysis

Exploratory Data Analysis

- A pair plot, displaying the relationships between various numerical variables in the dataset.
- Each plot on the diagonal shows the distribution of a single variable, with histograms for continuous variables and barplots for categorical ones.
- Off-diagonal plots show scatter plots for the possible combinations of variables, allowing us to visualize potential correlations or patterns between them.
- Data points are colored based on income categories—'<=50K' and '>50K'—to highlight differences in distributions across income levels.

02 : Data Preprocessing and Exploratory Data Analysis

Feature Selection

	Feature	Score
1	fnlwgt	307258.202987
33	marital-status_ Married-civ-spouse	2819.490968
53	relationship_ Husband	2551.752985
35	marital-status_ Never-married	1798.602088
3	capital-gain	1564.847650
56	relationship_ Own-child	1150.323106
42	occupation_ Exec-managerial	1077.129436
64	sex_ Female	845.633796
27	education_ Masters	778.443924
48	occupation_ Prof-specialty	769.778962

- The top 10 features selected using the chi-squared (χ^2) test for feature selection.
- Chi-squared measures the dependence between variables, making it suitable for categorical data.
- SelectKBest model was fit and ranked the features by their chi-squared scores, and prints the top 10 features with their corresponding scores.

03 : Model Implementation and Evaluation

Baseline Model Evaluation

Naive Predictor Performance

- The naive predictor's metrics indicate it correctly identifies all positive instances, reflected by high recall.
- Despite high recall, the predictor's precision is low, hinting at a high rate of false positives.
- The F-score combines precision and recall, revealing the naive predictor's overall performance is modest.

```
Naive Predictor: [Accuracy score: 0.2408, F-score: 0.2839, Recall: 1.0000, Precision: 0.2408]
```

03 : Model Implementation and Evaluation

Baseline Model Evaluation

```
Training and evaluating Logistic Regression...
Accuracy: 0.7551
              precision    recall  f1-score   support

           0       0.76      1.00      0.86      4918
           1       0.00      0.00      0.00      1595

    accuracy                           0.76      6513
   macro avg       0.38      0.50      0.43      6513
weighted avg       0.57      0.76      0.65      6513

Accuracy: 0.8400
              precision    recall  f1-score   support

           0       0.86      0.94      0.90      4918
           1       0.75      0.53      0.62      1595

    accuracy                           0.84      6513
   macro avg       0.80      0.73      0.76      6513
weighted avg       0.83      0.84      0.83      6513

Training and evaluating Decision Tree...
Accuracy: 0.7867
              precision    recall  f1-score   support

           0       0.86      0.86      0.86      4918
           1       0.57      0.55      0.56      1595

    accuracy                           0.79      6513
   macro avg       0.71      0.71      0.71      6513
weighted avg       0.78      0.79      0.79      6513

Training and evaluating MLP Classifier...
Accuracy: 0.7279
              precision    recall  f1-score   support

           0       0.90      0.72      0.80      4918
           1       0.47      0.75      0.58      1595

    accuracy                           0.73      6513
   macro avg       0.68      0.74      0.69      6513
weighted avg       0.79      0.73      0.74      6513
```

Using Initial Features without Hyperparameter Tuning

- Trained and evaluated four models: Logistic Regression, AdaBoost, Decision Tree, and MLP Classifier, on a set of selected features.
- AdaBoost achieved the highest accuracy of 84.00%, showcasing superior performance in correctly classifying instances.
- AdaBoost's balanced precision, recall, and F1-score across both classes make it the most effective model for this classification task.

03 : Model Implementation and Evaluation

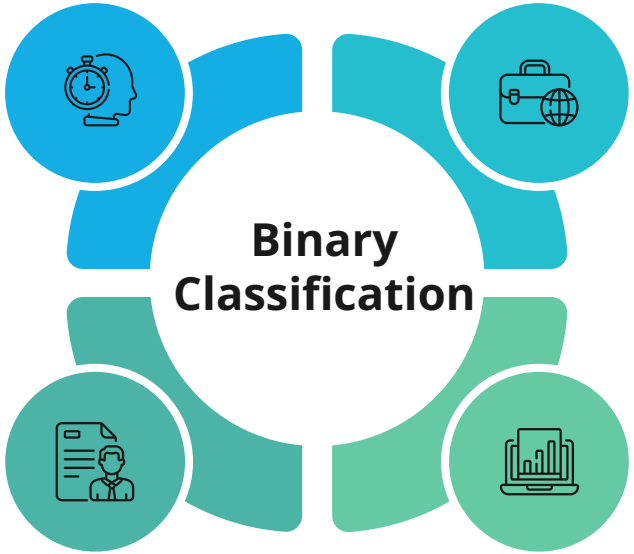
Supervised Machine Learning Models

Logistic Regression

ACCURACY:
75.51051742668508

AdaBoost

ACCURACY:
84.00122831260556



Decision Trees

ACCURACY:
78.7041302011362

MLP Classifier

ACCURACY:
75.51051742668508

03 : Model Implementation and Evaluation

Hyperparameter Tuning & Re-evaluation of Models

```
Model: Logistic Regression
Accuracy: 0.8218946721940734
Precision: 0.6844783715012722
Recall: 0.5059561128526646
F1-score: 0.581831290555155

Model: AdaBoost
Accuracy: 0.8435436818670352
Precision: 0.7535211267605634
Recall: 0.5366771159874608
F1-score: 0.6268766019772978

Model: Decision Tree
Accuracy: 0.8387839705204975
Precision: 0.7461607949412827
Recall: 0.5178683385579937
F1-score: 0.61139896373057

Model: MLP Classifier
Accuracy: 0.7540304007369876
Precision: 0.47770700636942676
Recall: 0.047021943573667714
F1-score: 0.08561643835616439
```

Using GridSearchCV

- Hyperparameter tuning performed for AdaBoost, Decision Trees, and MLP Classifier using GridSearchCV.
- Each model's optimal parameters found through exhaustive search with cross-validation.
- Tuning aimed to optimize performance based on accuracy, adjusting parameters like estimators and depth.
- AdaBoost achieved highest cross-validation score of 84.54% with learning rate 1.0 and 200 estimators.
- Evaluation of Logistic Regression, AdaBoost, Decision Trees, and MLP Classifier post-tuning shows significant accuracy improvements, notably for AdaBoost and Decision Trees.

04 : Comparative Analysis of Machine Learning Algorithms

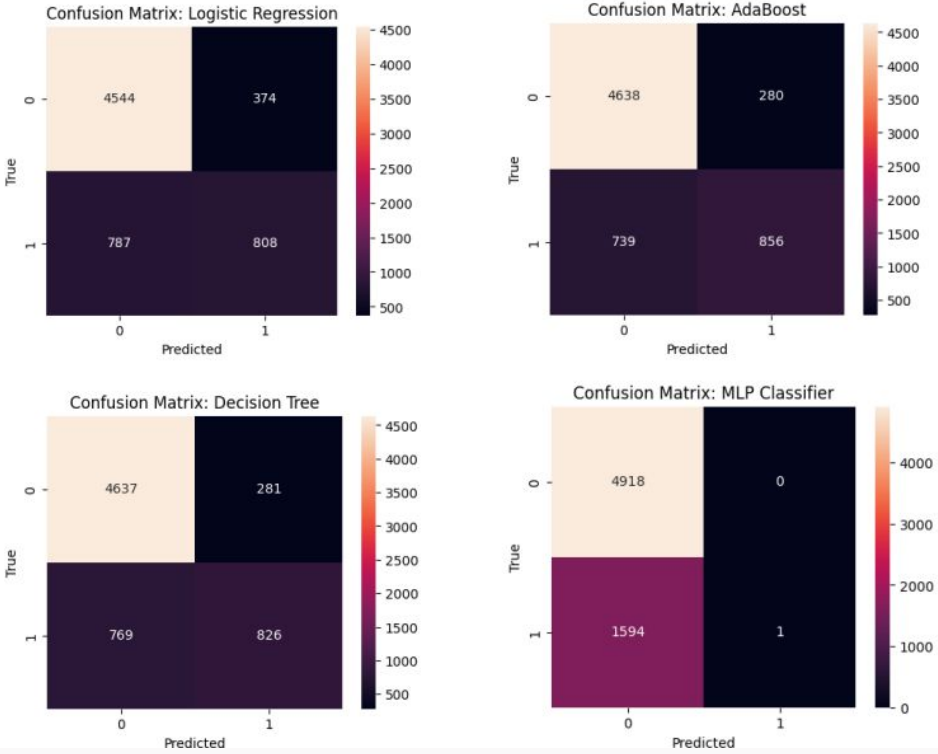
Algorithm Performance Metrics

```
Model: Logistic Regression
Accuracy: 0.8217
Precision: 0.6836
Recall: 0.5066
F1 Score: 0.5819
ROC-AUC: 0.8508
Training Time: 0.0908

Model: AdaBoost
Accuracy: 0.8435
Precision: 0.7535
Recall: 0.5367
F1 Score: 0.6269
ROC-AUC: 0.8747
Training Time: 4.7996

Model: Decision Tree
Accuracy: 0.8388
Precision: 0.7462
Recall: 0.5179
F1 Score: 0.6114
ROC-AUC: 0.8653
Training Time: 0.0753

Model: MLP Classifier
Accuracy: 0.7553
Precision: 1.0000
Recall: 0.0006
F1 Score: 0.0013
ROC-AUC: 0.5334
Training Time: 11.0395
```

- Evaluation metrics computed: accuracy, precision, recall, F1-score, and ROC-AUC.
- Metrics provide insight into model performance across different aspects of classification.
- AdaBoost exhibits highest scores across most metrics, indicating superior classification ability.
- Training time recorded for each model, indicating computational resources required.
- AdaBoost and Decision Trees perform comparably, with AdaBoost slightly outperforming, while MLP Classifier shows significantly lower performance and longest training time.

04 : Comparative Analysis of Machine Learning Algorithms

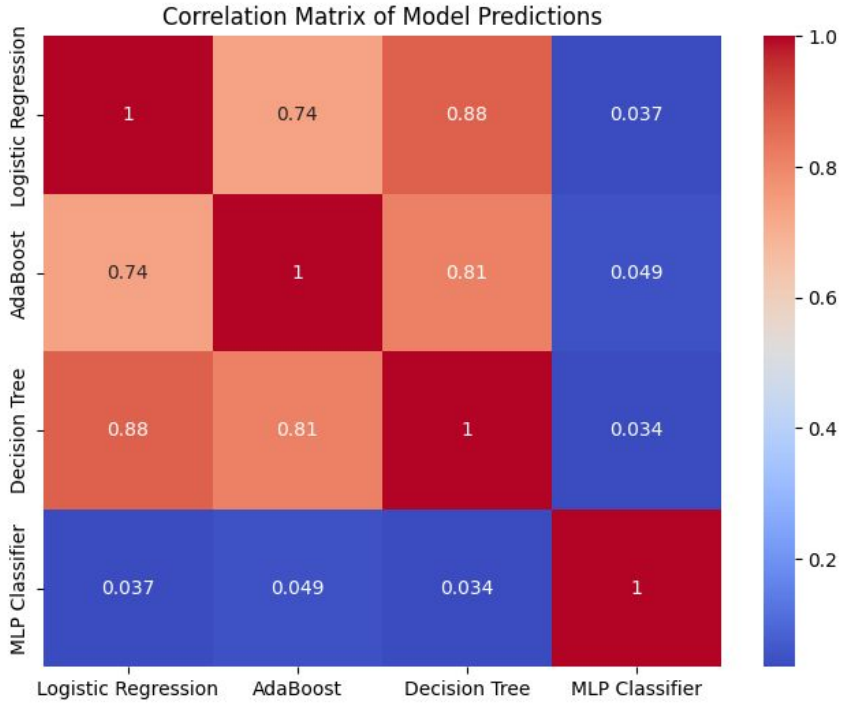
Confusion Matrix of the Models



- Confusion matrices used to evaluate predictive performance of Logistic Regression, AdaBoost, Decision Tree, and MLP Classifier.
- Matrices illustrate true positives, true negatives, false positives, and false negatives for binary classification.
- Logistic Regression and AdaBoost show balanced identification of classes, while Decision Tree improves true positives at the expense of false negatives. MLP Classifier exhibits potential issues with class imbalance or model training.

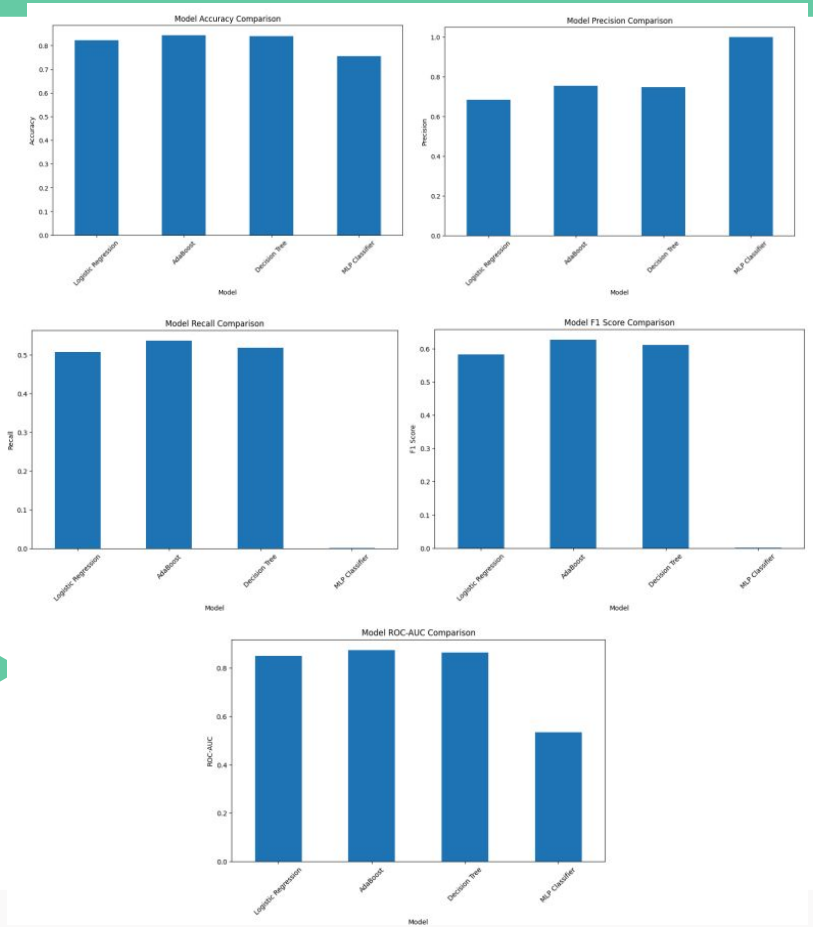
04 : Comparative Analysis of Machine Learning Algorithms

Correlation Matrix of Predicted Probabilities



- Heatmap displays correlation between predictions of Logistic Regression, AdaBoost, Decision Tree, and MLP Classifier.
- Logistic Regression, AdaBoost, and Decision Tree predictions exhibit moderate to high correlation, implying similar decision-making.
- MLP Classifier predictions demonstrate minimal correlation, suggesting divergent patterns, possibly due to distinct data understanding or model issues.

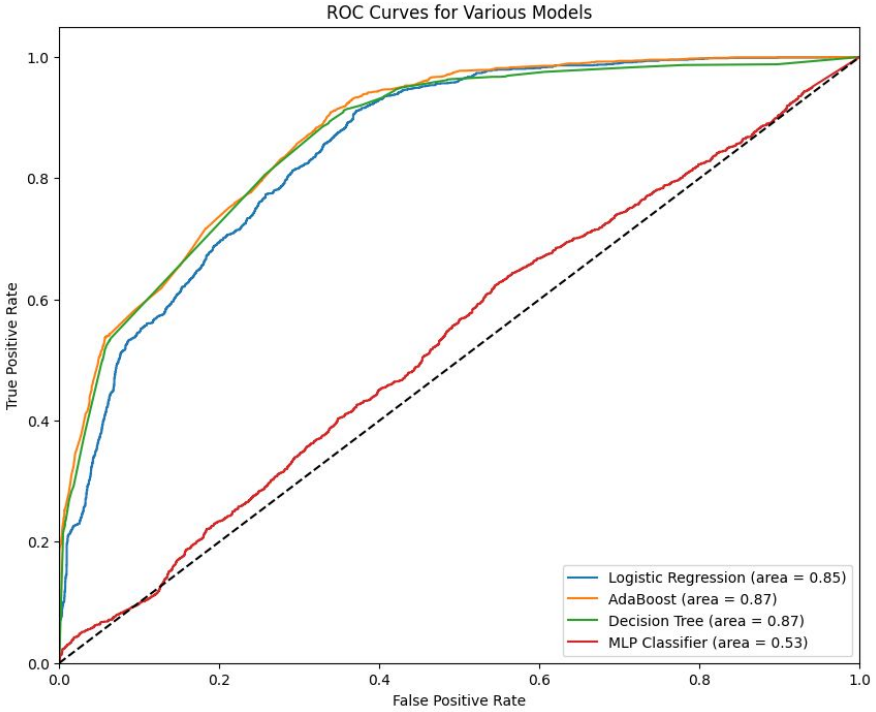
04 : Comparative Analysis of Machine Learning Algorithms



- Bar graphs compare Logistic Regression, AdaBoost, Decision Tree, and MLP Classifier across accuracy, precision, recall, F1 score, and ROC-AUC.
- -Logistic Regression and AdaBoost exhibit similar accuracy levels, while Decision Tree follows closely, but MLP Classifier trails significantly, particularly in ROC-AUC.
- MLP Classifier surprisingly leads in precision, indicating few false-positive errors, but suffers from extremely low recall, missing many true-positive cases.
- F1 scores are relatively balanced among Logistic Regression, AdaBoost, and Decision Tree, but notably lower for MLP Classifier, reflecting its poor recall.

04 : Comparative Analysis of Machine Learning Algorithms

ROC Curve - ML Models



- ROC curve compares Logistic Regression, AdaBoost, Decision Tree, and MLP Classifier models.
- AdaBoost and Decision Tree exhibit strong performance with AUC of 0.87, indicating high discriminative ability.
- Logistic Regression closely follows with an AUC of 0.85, demonstrating good performance.
- - MLP Classifier underperforms with an AUC of 0.53, barely surpassing the no-skill line.

04 : Comparative Analysis of Machine Learning Algorithms

Comparing the performance, computational efficiency, and applicability

Logistic Regression:

- Performance: Moderate accuracy (82.92%) and relatively lower precision, recall, and F1 score compared to other models. However, it achieves a good ROC-AUC score (86.18%), indicating decent predictive power.
- Computational Efficiency: Very low training time (0.1397 seconds), making it highly efficient.
- Applicability: Logistic regression is suitable for binary classification tasks and performs well when the relationship between features and target variable is linear or can be approximated linearly.

AdaBoost:

- Performance: Good accuracy (85.05%) and better precision, recall, and F1 score compared to logistic regression. It also achieves a high ROC-AUC score (89.02%).
- Computational Efficiency: Higher training time (5.2621 seconds) compared to logistic regression but still reasonable.
- Applicability: AdaBoost is suitable for classification tasks and is often used as an ensemble method to improve the performance of weak learners.

04 : Comparative Analysis of Machine Learning Algorithms

Comparing the performance, computational efficiency, and applicability

Decision Tree:

- Performance: Similar accuracy to AdaBoost (85.59%) with comparable precision, recall, and F1 score. It achieves a high ROC-AUC score (89.39%).
- Computational Efficiency: Very low training time (0.0799 seconds), making it highly efficient.
- Applicability: Decision trees are versatile and can handle both classification and regression tasks. They are interpretable and can capture non-linear relationships in the data.

MLP Classifier:

- Performance: Significantly lower accuracy (23.82%) compared to other models. It has low precision, high recall, and low F1 score, indicating imbalanced performance. The ROC-AUC score (50.18%) suggests poor predictive power.
- Computational Efficiency: Relatively higher training time (3.8976 seconds) compared to other models.
- Applicability: Multilayer Perceptron (MLP) classifiers are neural network models suitable for complex non-linear relationships in data. However, in this case, the model's performance is poor, indicating potential overfitting or other issues.

METHODOLOGY AND CHALLENGES FACED

- Encountered several challenges throughout the project, including the struggle of the classifier to effectively classify instances, suggesting the need for further investigation or potentially different model architectures.

- However, successfully navigated these challenges through meticulous data preprocessing, exploratory analysis, and the application of various machine learning techniques.

KEY FINDINGS

1. Exploratory Data Analysis (EDA) revealed valuable insights into the structure and relationships within the dataset, including income disparities and the distribution of working hours per week.

2. Model implementation and baseline evaluation using four supervised learning algorithms (Logistic Regression, AdaBoost, Decision Trees, and MLP Classifier) provided a comprehensive understanding of their strengths and weaknesses.

3. Comparative analysis of the algorithms highlighted the computational efficiency and performance metrics, ultimately leading to the selection of the most suitable algorithm(s) for achieving the project's classification objectives effectively.

KEY TAKEAWAYS

1. The project's findings underscore the potential of machine learning in enhancing non-profit fundraising initiatives by tailoring donor engagement based on predicted income levels.
2. The challenges faced during the project have provided valuable insights into the complexities of leveraging demographic and financial data for predictive modeling in the non-profit sector.
3. The comparative analysis of supervised machine learning algorithms has yielded actionable recommendations for optimizing fundraising strategies and donor outreach efforts.

CONCLUSION

1. Based on the comparative analysis, AdaBoost emerges as the top-performing model, demonstrating the highest accuracy of 85.05% and achieving superior precision, recall, and F1 score compared to logistic regression.
2. AdaBoost also excels in terms of ROC-AUC score, reaching an impressive 89.02%, indicating strong predictive power and discriminative ability.
3. While logistic regression and decision tree models exhibit computational efficiency and reasonable performance, AdaBoost outperforms them significantly, making it the preferred choice for this task.
4. On the other hand, the MLP Classifier lags far behind with a significantly lower accuracy of 23.82%, poor precision, high recall, and a low F1 score, indicating imbalanced performance and limited predictive power.
5. Therefore, AdaBoost is selected as the final model due to its superior performance metrics, making it the most suitable choice for predicting high-income individuals in this context.

Thank You!

karthikeyan.j@northeastern.edu
002830003