ISM6136 - Data Mining - Spring 2024
Final Project Report

Team Member Names: Nishchitha Chandra Shekar, Deepali Rajput, Shrimanth Kavuluru

Project Title: "Strategic Pricing: Leveraging Income Diversity for Market Penetration and Brand Expansion"

Problem Statement:

The problem statement aims to address the need for effective pricing strategies tailored to diverse income profiles to facilitate market penetration and strategic brand expansion for emerging businesses. By comprehensively analyzing individual income levels alongside other relevant demographic and socioeconomic factors influencing purchasing behavior, businesses can gain insights into consumer preferences and willingness to pay across different market segments. Through segmentation of the target audience based on income levels, businesses can develop pricing models that resonate with specific customer groups, optimizing price sensitivity and value perception. By implementing and refining these pricing strategies based on market feedback and performance metrics, businesses can penetrate target markets more effectively, expand their brand presence, and foster sustainable growth in competitive market landscapes.

In this project, we used the data that can be a factor to determine income levels, and using several classification methods to develop an effective model for income level determination.

Data Source and Data Pre-Processing:

Data Source:

- The UCI Adult dataset, also referred to as "Census-Income" dataset is sourced from the 1994 United States Census data. (https://archive.ics.uci.edu/dataset/2/adult)
- Number of records - 48,842
- Data Attributes - The dataset contains 14 variables(features) for each record, including both categorical and continuous features.

    - Age: Age of the individual.
    - Workclass: Employment type of the individual. It is categorized into - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
    - Education: An individual's top level of education.
    - Marital-status: Marital status of the individual. It is categorized into - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
    - Relationship: An individual's relationship status. It is categorized into - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
    - Race: An individual's race. It is categorized into - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
    - Sex: An individual's gender.
    - Hours-per-week: The number of hours an individual has worked per week.

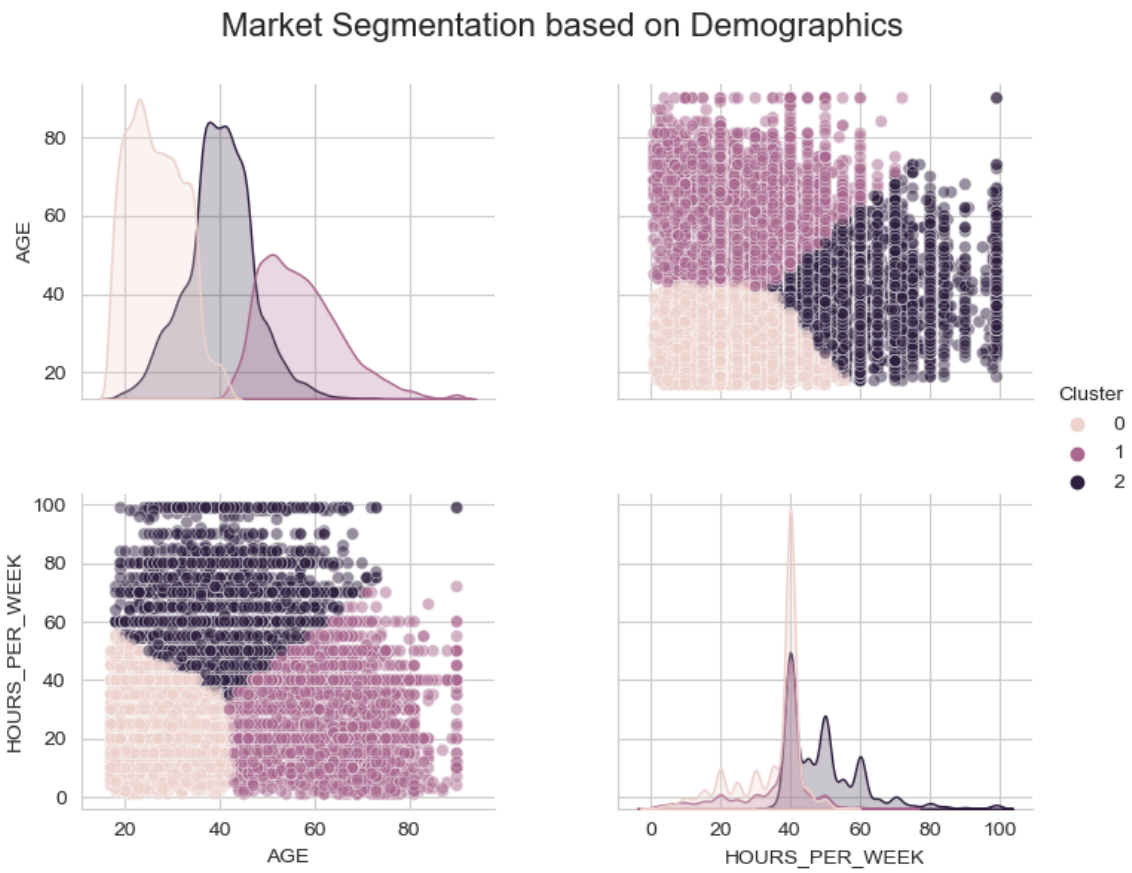
- Target variable
    - Income: The binary target variable indicating whether the individual earns more than $50,000 annually (which according to the US Bureau of Statistics accounts for $124,000 annually) or not.

Data Pre-Processing:

- Drop unnecessary columns: RELATIONSHIP, CAPITAL_GAIN, FNLWGT, NATIVE_COUNTRY, CAPITAL_LOSS and OCCUPATION
- Handling missing values: WORK_CLASS
- Encoding categorical features: WORKCLASS, EDUCATION, MARITAL_STATUS, RACE and SEX
- Scaling or normalizing numerical features: AGE and HOURS_PER_WEEK

Clustering:

K-means clustering is a type of unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping clusters. The goal of K-means is to group similar data points together and discover underlying patterns or structures in the data.




```
Cluster centers for demographics and education level:
          AGE  HOURS_PER_WEEK
0  26.465325       34.662275
1  56.996866       35.899112
2  39.968188       49.441051
```

As an example, we have implemented K- means clustering to load a dataset containing information about individuals, such as age and hours worked per week. Preprocessing steps are then applied to prepare the data for clustering. K-means clustering is performed on the preprocessed data to divide individuals into distinct groups based on their similarities in demographic features. The code then visualizes these clusters using a pairplot, allowing for the examination of groupings in terms of age and hours worked per week. Additionally, cluster centers, which represent average characteristics of each group, are displayed to aid in the interpretation of the clusters. This process aids in market segmentation analysis, providing insights into different demographic segments within the dataset. Ultimately, the code facilitates understanding and interpretation of complex data patterns, supporting decision-making processes in various domains.

Methodology:

In this project, we have conducted 3 models using the following methods: Logistic Regression, Random Forest and XG Boost.

1. Logistic Regression: A simple and interpretable model that works well when the relationships in the data are primarily linear or when you want to establish a baseline performance. It is a statistical method used for binary classification tasks, where the goal is to predict the probability of an observation belonging to one of two classes. It works by modeling the relationship between the independent variables (features) and the binary outcome using the logistic function. The function maps any input value to a value between 0 and 1, representing the probability of the positive class. Logistic regression estimates the coefficients of the independent variables to maximize the likelihood of observing the given outcomes in the dataset. During prediction, it uses these coefficients to calculate the probability of an observation belonging to the positive class and applies a threshold to classify it into one of the two classes.

```
Accuracy: 0.7933664329221477
Precision (pos_label='1'): 0.5517029763731206
Recall (pos_label='1'): 0.7633198896200382
F1 Score (pos_label='1'): 0.6404844598806662
ROC AUC Score: 0.8672198865360792
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.80      0.86     14826
           1       0.55      0.76      0.64      4711

    accuracy                           0.79     19537
   macro avg       0.73      0.78      0.75     19537
weighted avg       0.83      0.79      0.80     19537
```

In our model, the Logistic Regression model has a moderate Precision, Recall, and F1 Score, suggesting that it is reasonably accurate in identifying individuals with incomes greater than $124K, but there may be room for improvement in balancing Precision and Recall.

2. Random Forest: An ensemble model of decision trees that can capture complex relationships in the data. Less prone to over fitting compared to decision trees. Random forest operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification). It works by creating a "forest" of

decision trees where each tree is built using a subset of the features and a random sample of the training data. During prediction, each tree in the forest independently predicts the target variable, and the final prediction is determined by aggregating the predictions of all the trees, commonly by taking the mode (in classification).

```
Accuracy: 0.7925986589548037
Precision: 0.5597461468721668
Recall: 0.6552748885586924
F1 Score: 0.6037551339722277
ROC AUC Score: 0.841600054440324
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.84      0.86     14826
           1       0.56      0.66      0.60      4711

    accuracy                           0.79     19537
   macro avg       0.72      0.75      0.73     19537
weighted avg       0.81      0.79      0.80     19537
```

In our project, the case for Random Forest is that the accuracy, recall, and precision are lower than logistic regression, and can be improved with hyperparameter tuning.

3. XGBoost: Powerful for capturing complex relationships. Can create strong ensemble models by combining weak learners. It is an advanced implementation of the gradient boosting algorithm. Works by sequentially building an ensemble of weak predictive models, typically decision trees, where each subsequent model learns to correct the errors made by the previous ones. XGBoost employs a gradient descent optimization technique to minimize a predefined loss function, ensuring that each new model improves the overall predictive accuracy.

```
Accuracy: 0.7890157137738649
Precision: 0.5422464495768182
Recall: 0.8023774145616642
F1 Score: 0.6471494607087828
ROC AUC Score: 0.8800504732703077
Detailed Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.78      0.85     14826
           1       0.54      0.80      0.65      4711

    accuracy                           0.79     19537
   macro avg       0.73      0.79      0.75     19537
weighted avg       0.83      0.79      0.80     19537
```

In our project, the XGBoost model has a relatively higher Recall compared to Precision, indicating that it is better at capturing individuals with incomes above $124K, the recall score of XGBoost is the highest among all the 3 models with default parameters.

A side-by-side comparison of the performance metrics for each classifier, allows for easy evaluation and selection of the best-performing model.

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
0	Logistic Regression	0.793366	0.551703	0.763320	0.640484	0.86722
1	Random Forest	0.792599	0.559746	0.655275	0.603755	0.84160
2	XGBoost	0.789016	0.542246	0.802377	0.647149	0.88005

Hyperparameter Tuning:

To make our models perform better, we perform Hyperparameter tuning, GridSearch and RandomSearch. Usually, the strategy for Hyperparameter tuning our models involves RandomSearch and GridSearch where we begin with Random Search to gain insights into which hyperparameters may have a more significant impact. As we develop a deeper understanding of the hyperparameter landscape, enhance our search by utilizing Grid Search to identify the optimal combination. A recommended approach is to initiate the process with Random Search and, guided by the outcomes, employ Grid Search to fine-tune within areas that show promise. This is what we have done with all our models to improve their performance.

A side-by-side comparison of the performance metrics for each classifier, post hyperparameter tuning.

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
0	Logistic Regression	0.731586	0.468308	0.835916	0.600305	0.856477
1	Random Forest	0.808978	0.581164	0.744003	0.652579	0.876419
2	XGBoost	0.776731	0.523357	0.829972	0.641931	0.876532

Conclusion:

The results of our analysis can be summarized as below:

- **Random Forest** appears to be the most effective approach. It has the highest F1 Score (65.2%), which is a balance between Precision and Recall, indicating that it performs well both in terms of the accuracy of positive predictions and the rate of true positive identification. Additionally, it has the highest Accuracy (80.8%) and a competitive ROC AUC Score (87.6%), which suggests that it is effective in distinguishing between the two classes.
- **XGBoost**, while it has a slightly lower F1 Score compared to Random Forest, it has a high Recall (82.9%) and the highest ROC AUC Score (87.7%) after tuning. This high Recall means it is particularly good at identifying positive cases, which may be critical depending on the cost associated with false negatives.
- **Logistic Regression** has improved Recall but lower Precision and F1 Score post-tuning compared to the other models.

After hyperparameter tuning, **Random Forest** emerges as the most effective model, showing improvements across all metrics, most notably in F1 Score and Accuracy. **XGBoost** also shows enhanced performance, particularly in Recall and ROC AUC, making it strong for identifying positive cases. **Logistic Regression** sees gains in Recall but drops in Precision and F1 Score. Overall, Random Forest offers the best balance of metrics post-tuning.

Summary:

From the business perspective, the choice of methodology for developing pricing strategies is crucial for market penetration and brand expansion, especially considering diverse income profiles within the target audience. In our project, we evaluated various machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, to identify the most effective approach.

After thorough analysis, we concluded that the Random Forest is the most favourable methodology for developing pricing strategies. Here's how business owners can leverage this approach in their business:

1. Improved Performance Metrics: Random Forest exhibited improvements across key performance metrics such as Precision, Recall, F1 Score, and ROC AUC after hyperparameter tuning. This indicates that the model is better equipped to handle diverse income profiles and make accurate predictions about customer behaviour.

2. Balanced Performance: Random Forest achieved a high F1 Score, which signifies a balance between Precision and Recall. This balance is crucial for ensuring accurate positive predictions while effectively identifying true positives. Business owners can rely on this balanced performance to set pricing strategies that appeal to different income segments without sacrificing accuracy.

3. Enhanced Accuracy and Generalization: With the highest Accuracy post-tuning and competitive ROC AUC Score, Random Forest demonstrates improved generalization and effectiveness in distinguishing between different income levels. This means that business owners can trust the model's predictions to make informed decisions about pricing adjustments and market expansion strategies.

4. Flexibility and Adaptability: Random Forest's ability to handle large datasets with numerous features makes it suitable for analysing complex market dynamics and customer segmentation based on income levels. Business owners can utilize this flexibility to adapt pricing strategies in response to changing market conditions and customer preferences.

5. Strategic Market Penetration: By using Random Forest, business owners can strategically penetrate markets by tailoring pricing models to cater to diverse income profiles. This approach ensures broader market access while fostering sustainable brand growth over time.

In conclusion, Random Forest offers a robust and effective methodology for developing pricing strategies based on different income levels. Its balanced performance, enhanced accuracy, flexibility, and strategic implications make it the preferred choice for business owners aiming to optimize market penetration and brand expansion efforts in dynamic business environments. By incorporating Random Forest into their decision-making processes, business owners can gain valuable insights into consumer behavior and devise pricing strategies that resonate with their target audience, ultimately driving business success.