Department of Management Sciences - IIT Kanpur

DMS 672 – Data Mining

Project Report

Group 3 - Bias in Advertising Data

Submitted to:
Dr. Faiz Hamid
Associate Professor
DOMS IIT Kanpur

Name	Roll Number
Akash Tripathi	231250015
Anshi Srivastav	241140602
Devaguptapu Rama Krishna Sandilya	231250043
Varun Pratap Singh Chauhan	231250161

Table Of Contents

1. Introduction

Overview of the problem statement

Classification of users/customers conversion based on an advertisement shown.

Methodology

Based on the available dataset, classification models to classify the true conversion of the customer to be built. Prior to building the models, EDA and pre-processing are done to clean the data. Hence, the project is divided into 2 phases:

- Phase 1: EDA and Data Preprocessing, and
- Phase 2: Application of DM algorithm and Result analysis

Brief description of the dataset and objectives:

Dataset contains synthetic generated data for users who were shown a certain advertisement (ad). Each instance of the dataset is specific to a user and has feature attributes such as gender, age, income, political/religious affiliation, parental status, home ownership, area (rural/urban), and education status.

In addition to the features, it also contains information on whether users actually clicked on or were predicted to click on the ad. Clicking on the ad is known as conversion, and the three outcome variables included are: (1) The predicted probability of conversion, (2) Predicted conversion (binary 0/1) which is obtained by thresholding the predicted probability, (3) True conversion (binary 0/1) that indicates whether the user actually clicked on the ad.

2. Exploratory Data Analysis (EDA) and Data Preprocessing

Description of the features in the Dataset

- Age
- Gender
- Income
- Political / Religious Affiliation
- Parental Status
- Home Ownership
- Area (Rural/Urban)
- Education Status

Target Variable in the Dataset – Conversion

- A user is considered to have converted (true conversion=1) if they clicked on the ad.
- *Predicted_Conversion, True_Conversion* (Estimated, Actual) data given
- *Predicted_Conversion* is obtained by thresholding the predicted probability, provided in the dataset

Dataset Overview:

The dataset comprises demographic variables like age, gender, income, education, and politics, alongside conversion metrics (*true_conversion*, *predicted_conversion*). The goal is to uncover patterns that may highlight potential biases in targeted ad delivery. Below are key features:

- **Demographics**: age, gender, income, politics, religion, college_educated, parents, homeowner, area.
- **Conversion Indicators**: true_conversion (actual conversions) and predicted_conversion (conversion estimate from the model).
- **Additional Metrics**: predicted_probability, the model's probability output.

Data Inspection:

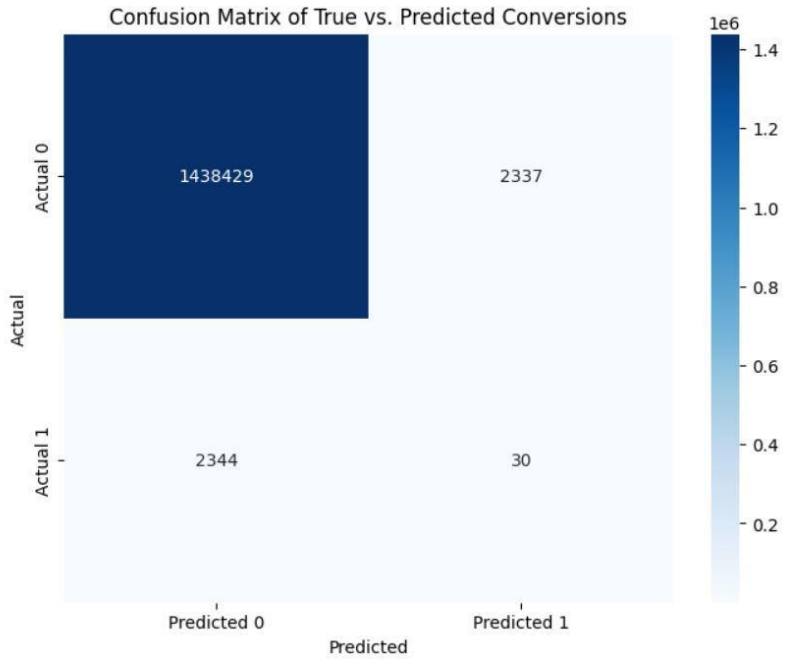
Initial EDA involves identifying unique values, understanding data distribution, and flagging any potential quality issues:

- **Unique Values Analysis**: We observed that features like gender, religion, and politics contain a significant number of Unknown values, signalling potential data gaps that could bias the analysis if not addressed.
- **Summary Statistics:** Mean, median, and mode statistics were calculated for continuous features like *predicted_probability* and income.
- **Missing Data Patterns:** Features with frequent Unknown values were assessed for imputation or exclusion.

EDA / Preprocessing Steps:

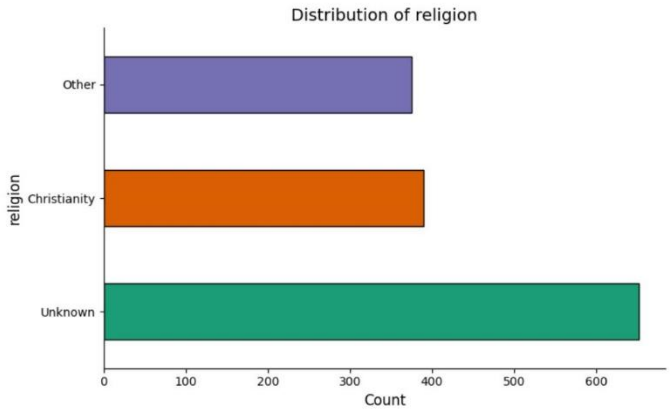
Steps

1. Extracting Unique Values in each of the Features of the Dataset and finding the number of Unknown / Missing values in each column.
2. Finding Occurrences / Data Points where the Conversion differs despite being the same on all other features and comparing True vs. Predicted Conversion from the Dataset.

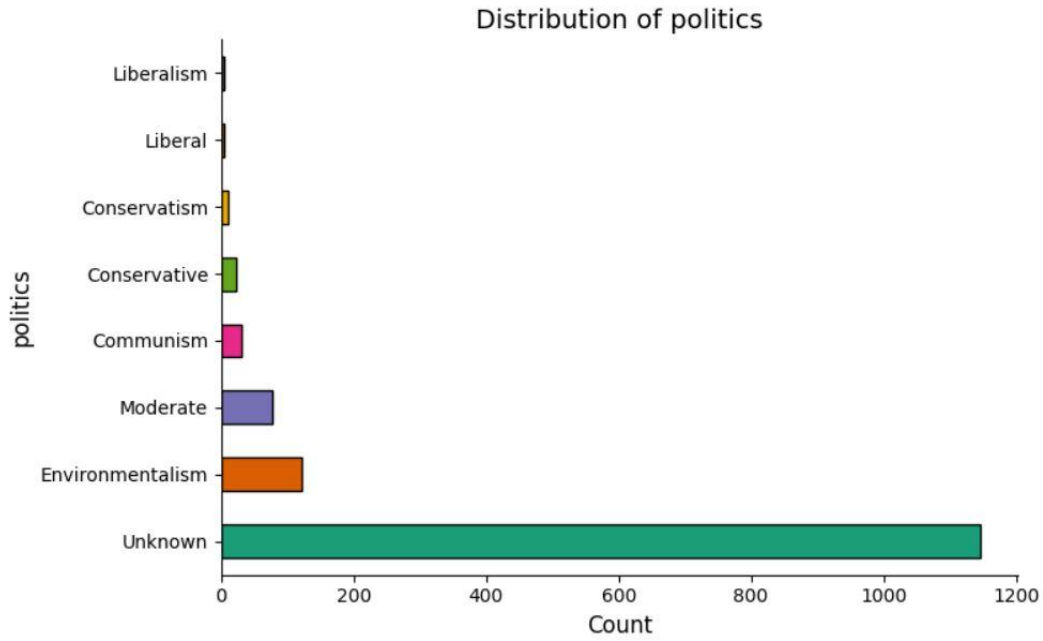


3. Finding the threshold for *predicted_probability* where *predicted_conversion* is 1 happens to be 0.3658223221090007
4. We dropped the column of *predicted_probability*
5. Also dropped the duplicate rows in the original dataset. (1443140 rows reduced to 2591 rows)

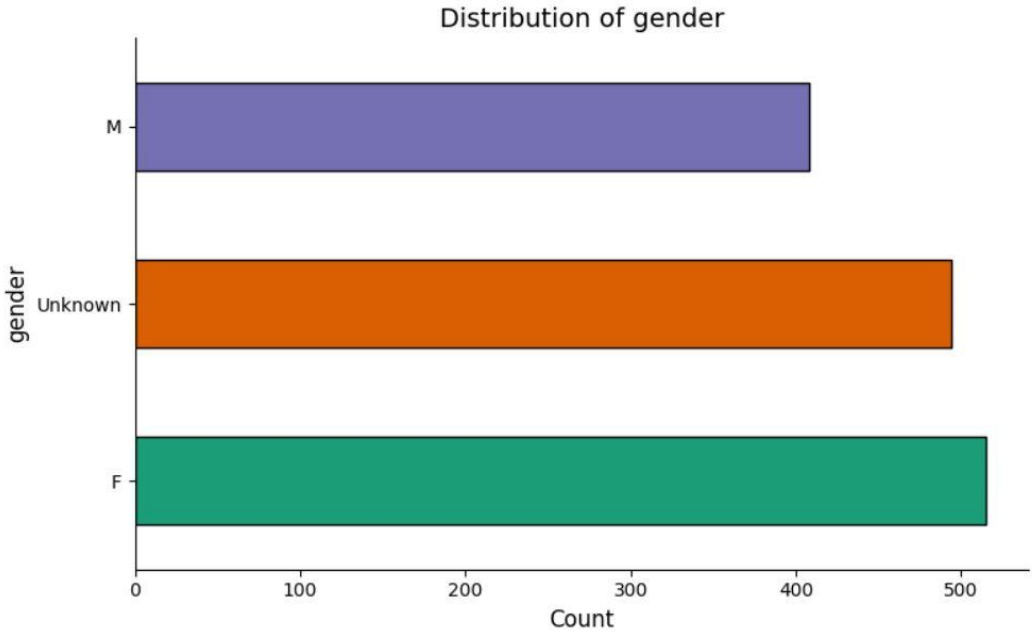
Univariate Analysis



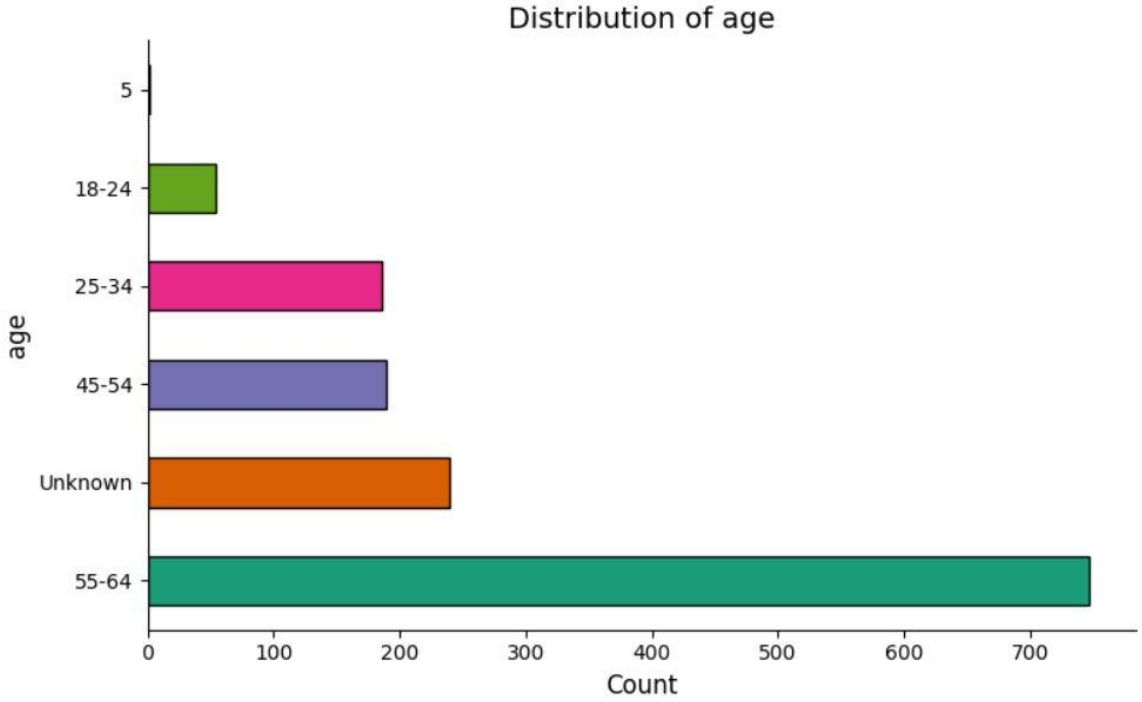
Religion: The data shows a few main categories, with "Christianity" being prominent among the identifiable groups, though a significant portion is labelled "Unknown." This could indicate missing data or people choosing not to specify their religious affiliation.



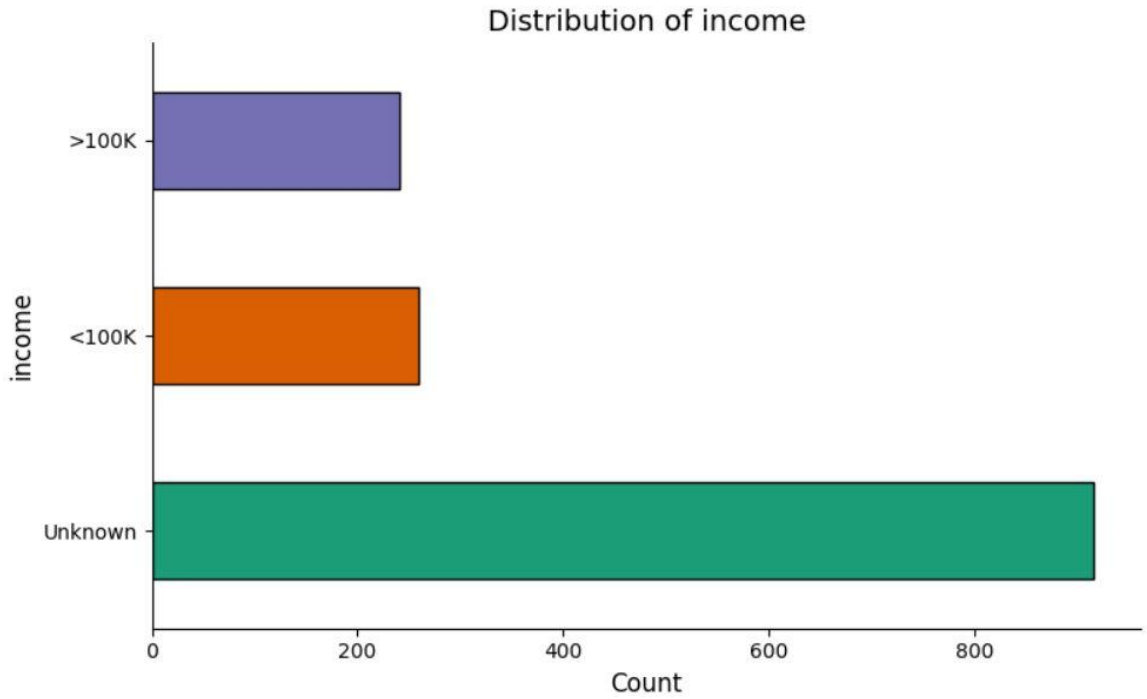
Politics: The "Politics" distribution is diverse, showing a range of political affiliations. Notably, there's a substantial "Unknown" category, and several distinct groups such as "Liberal," "Conservative," and "Moderate." This distribution suggests a varied set of political beliefs in the dataset.



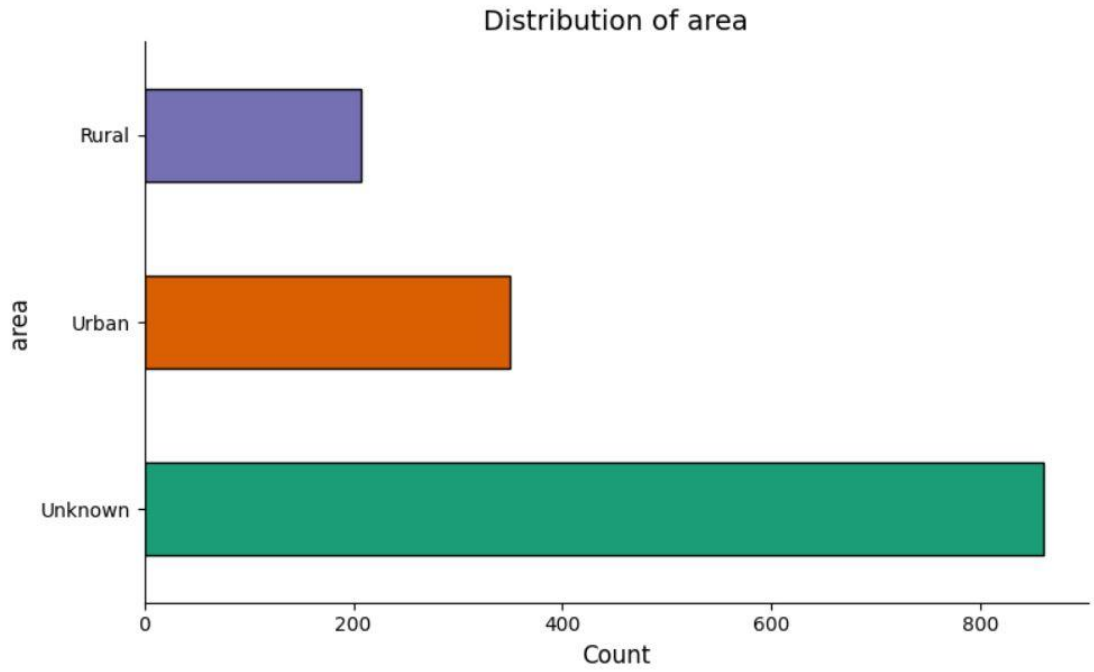
Gender: The "Gender" distribution is dominated by two main categories, with "M" and "F" labels, along with a considerable "Unknown" group. This may reflect non-disclosure or additional categories that weren't specified.



Age: The age distribution shows multiple age ranges, with certain ranges, such as "18-24" and "45-54," being more prominent. The "Unknown" category is also present here, hinting at either non-responses or unreported age data.

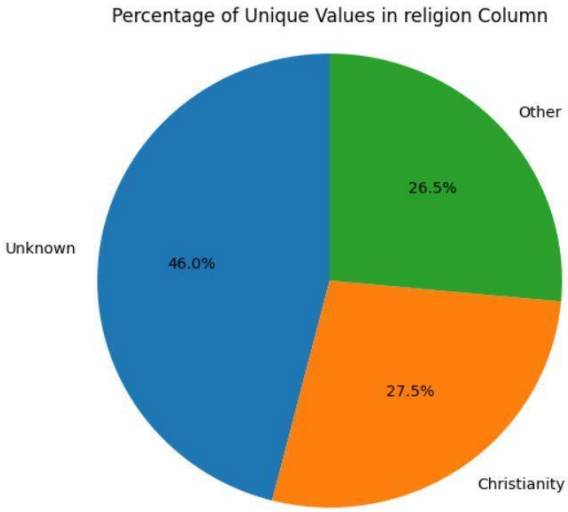


Income: The income distribution plot suggests that income levels are either grouped into broad ranges ("<100K" and ">100K") or marked as "Unknown." This wide range suggest a variety of socioeconomic backgrounds among the individuals in the dataset.

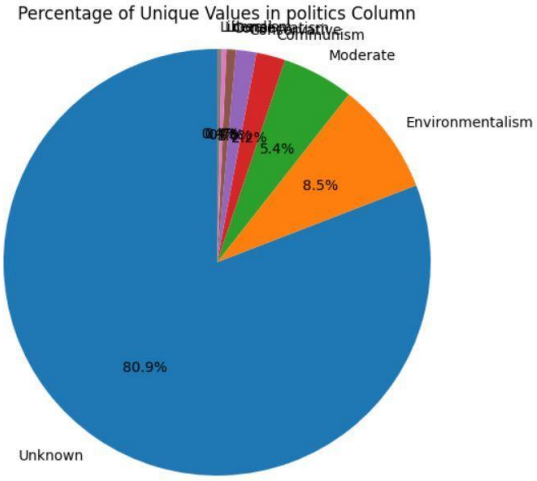


Area: The "Area" distribution reveals whether individuals are from "Urban" or "Rural" settings. There's also an "Unknown" category, which suggests that some location data was not disclosed.

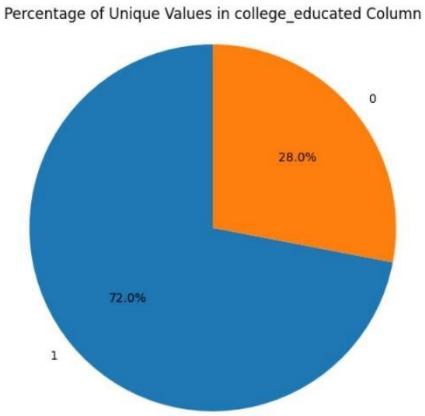
Unique values distribution

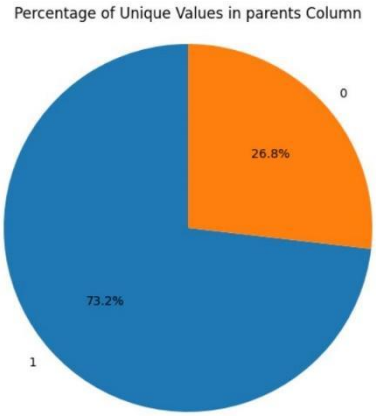


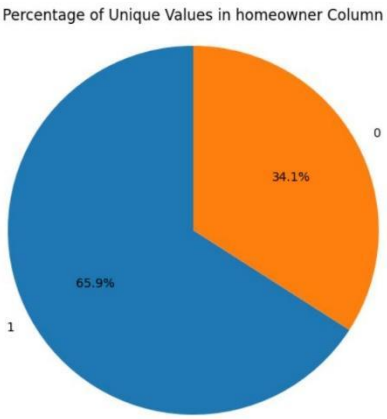
Religion: A substantial portion (46%) of the data is labelled as "Unknown," with "Christianity" and "Other" making up the rest. This suggests missing or unspecified data for a large segment of entries.



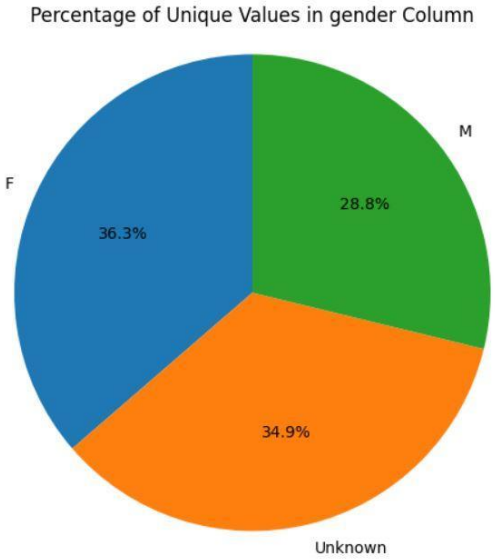
Politics: The "Politics" column shows an even larger "Unknown" segment (80.9%), with small percentages distributed among diverse political affiliations such as "Environmentalism," "Moderate," and "Communism." This high level of unspecified data might limit insights into political orientation.



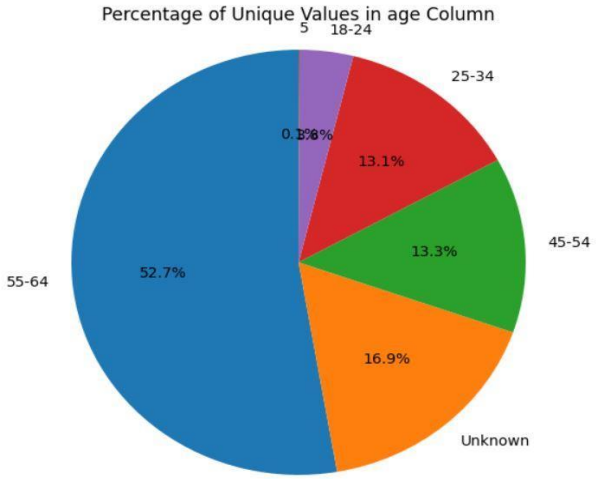




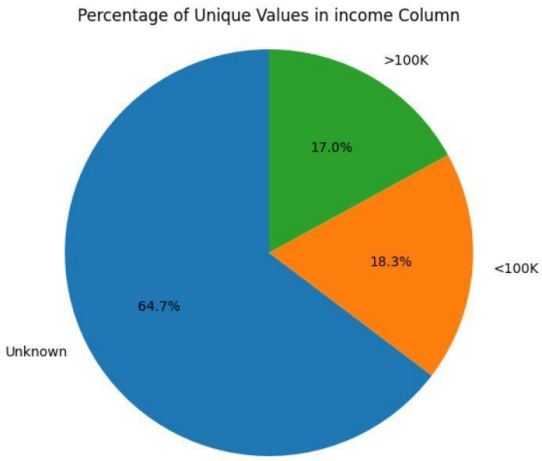
College Educated, Parents, and Homeowner Columns: These columns have a binary distribution, with most values being "1" (indicating "Yes" or "True"). This suggests that a majority of individuals are college-educated, parents, and homeowners.



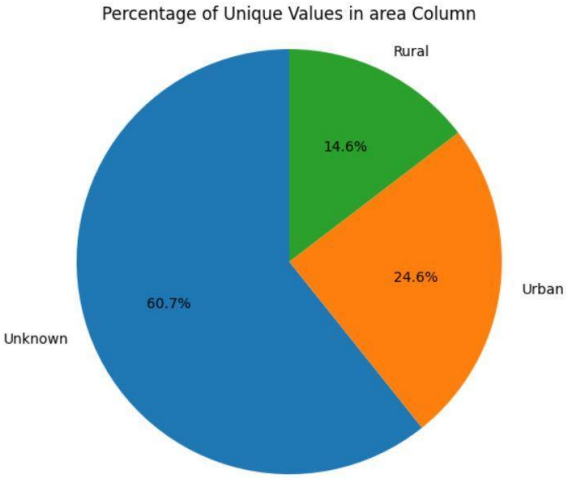
Gender: The gender distribution is fairly balanced, though slightly skewed, with "F" (36.3%) and "M" (28.8%). There's also a significant "Unknown" category (34.9%), which may reflect non-disclosure or inclusivity of unspecified gender options.



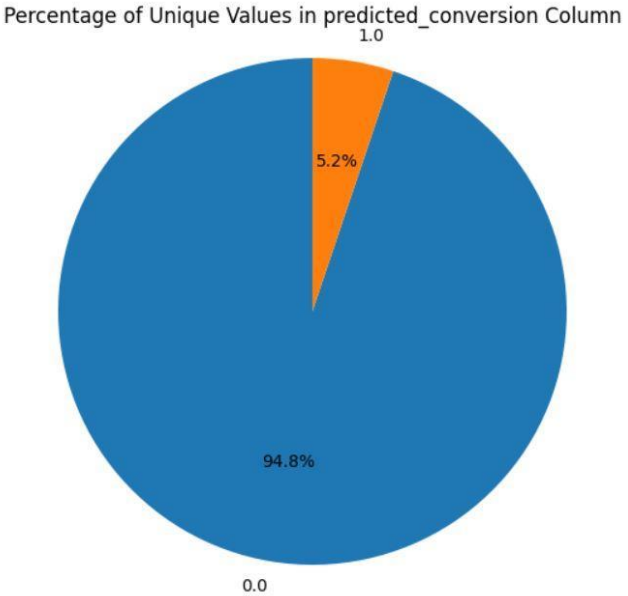
Age: The age distribution shows that most respondents are in the "55-64" range (52.7%), with a smaller percentage spread across other age groups and a notable "Unknown" segment (16.9%).

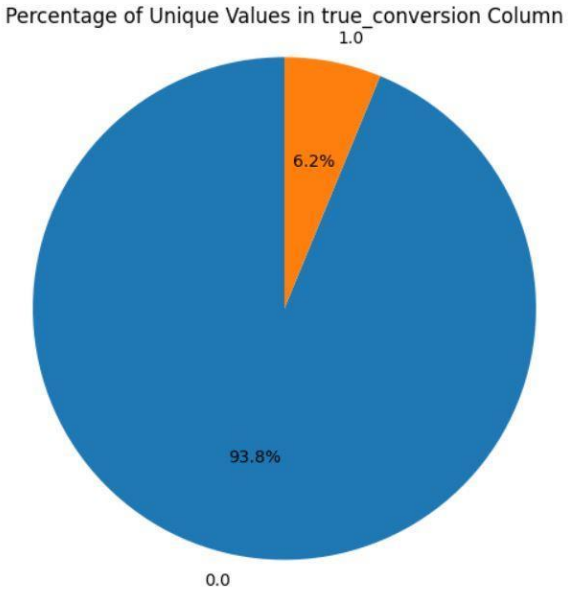


Income: Income data is also marked by a high "Unknown" value (64.7%), with the remainder fairly evenly split between "<100K" and ">100K."



Area: The area distribution shows a predominance of "Unknown" values (60.7%), followed by "Urban" and "Rural" groups.





Conversion Columns: Both "true_conversion" and "predicted_conversion" are predominantly "0.0," indicating low actual and predicted conversions in the dataset.

Handling Missing Data

As the missing data is very high, instead of dropping the column, we are replacing the values of "Unknown" in the dataset based on the proportion of other unique values in that specific column / feature of the dataset

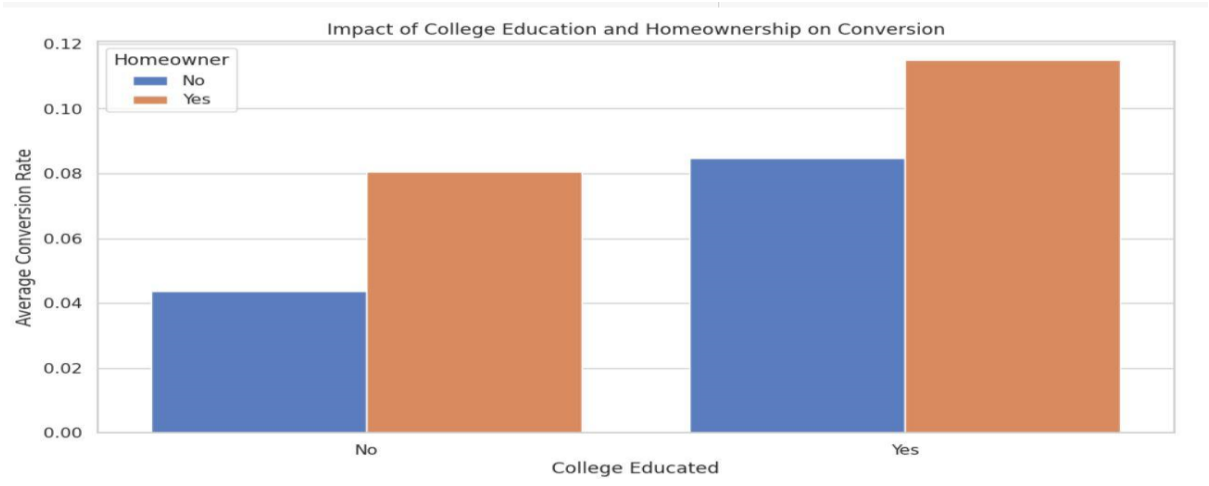
```
Count of 'Unknown' values filled in each column:
religion: 1147 'Unknown' values filled
politics: 1863 'Unknown' values filled
gender: 899 'Unknown' values filled
age: 436 'Unknown' values filled
income: 1621 'Unknown' values filled
area: 1492 'Unknown' values filled

Final count of 'Unknown' values in each column after filling:
religion              0
politics              0
college_educated      0
parents               0
homeowner             0
gender                0
age                   0
income                0
area                  0
true_conversion       0
predicted_conversion  0
dtype: int64
```

	religion	politics	college_educated	parents	homeowner	gender	age	income	area	true_conversion	predicted_conversion
0	Other	Environmentalism	1	1	1	M	55-64	<100K	Rural	0	0
1	Other	Communism	1	1	1	M	55-64	>100K	Urban	0	0
2	Other	Moderate	1	1	1	F	55-64	>100K	Urban	0	0
4	Other	Moderate	1	1	1	F	55-64	>100K	Urban	0	0
6	Other	Environmentalism	1	1	1	M	55-64	>100K	Rural	0	0

Bivariate Analysis

Understanding the impact of two features on the Predicted Conversion rate

1. **Impact of College Education & Home Ownership on Conversion**

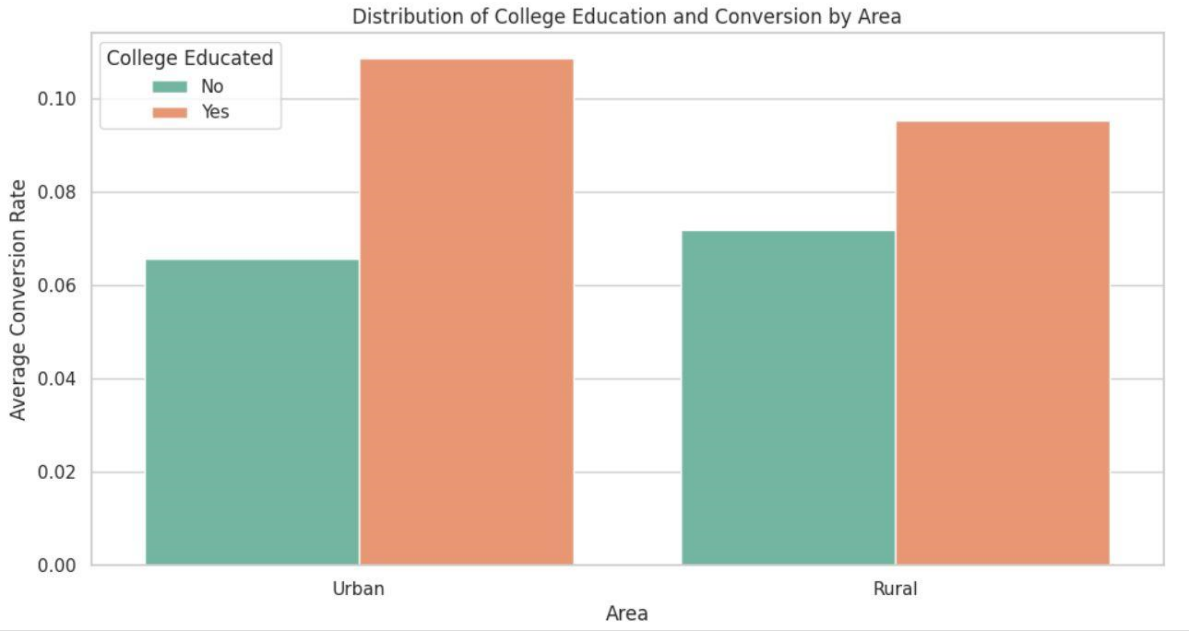


The bar chart shows the impact of college education and homeownership on conversion rates. It compares conversion rates across four groups based on college education status (Yes/No) and homeownership status (Yes/No).

- **Homeownership Effect**: Across both education levels, homeowners (orange bars) have higher conversion rates compared to non-homeowners (blue bars).
- **Education Effect**: People with a college education (right side) have higher conversion rates than those without a college education (left side).
- **Combined Impact**: The highest conversion rate is observed among college-educated homeowners, while the lowest conversion rate is among non-college-educated non-homeowners.

This suggests that both college education and homeownership positively impact conversion rates, with a more pronounced effect when combined.

2. **Impact of College Education & Area on Conversion**



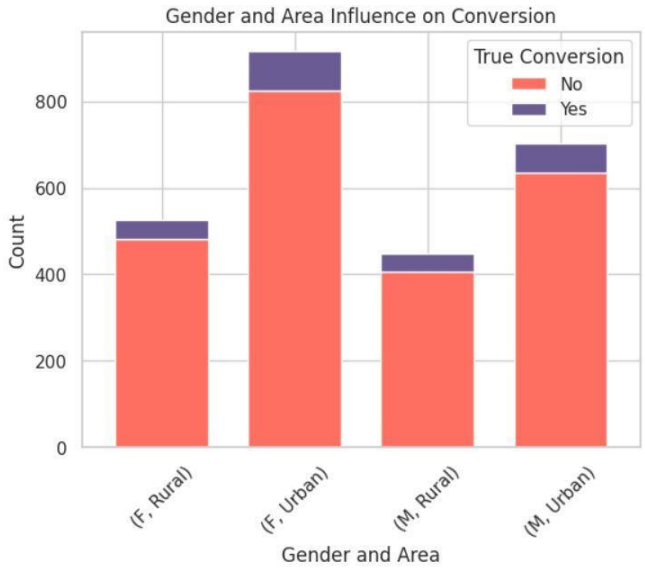
The bar chart presents the distribution of college education and conversion rates across different areas (Urban and Rural). It compares conversion rates based on whether individuals are college-educated (Yes/No)

- **Education Effect:** In both urban and rural areas, those with a college education (orange bars) have higher conversion rates compared to those without a college education (green bars).
- **Area Effect:** There is a similar trend in both urban and rural areas, with college-educated individuals converting at a higher rate than non-college-educated individuals.

- **High Conversion Rates:** College-educated individuals in both areas have notably higher conversion rates, suggesting that education level significantly influences conversions regardless of the area.

This indicates that college education positively impacts conversion rates in both urban and rural settings.

3. **Impact of Gender & Area on Conversion**



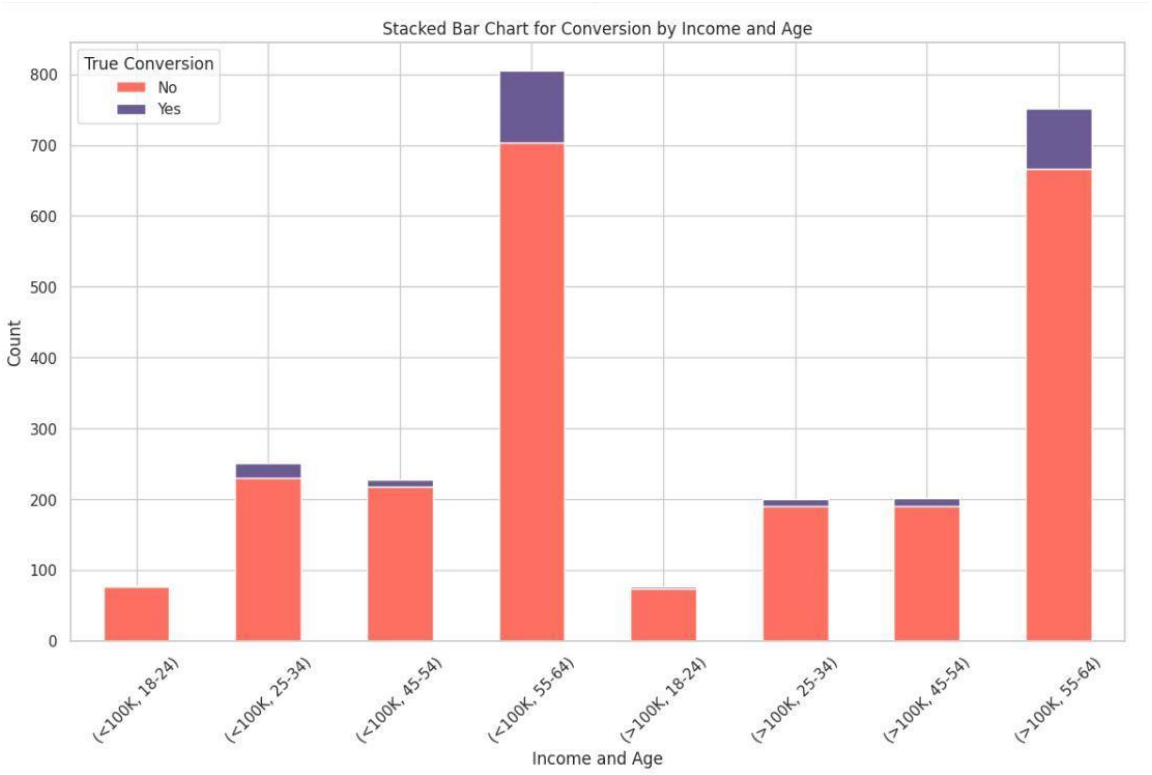
This stacked bar chart illustrates the influence of gender and area (urban or rural) on conversion rates, with separate counts for individuals who converted (Yes) and those who did not (No).

Key observations:

1. **Gender and Area Breakdown**:
    - Urban females ("F, Urban") have the highest count, but the proportion of conversions (Yes) is relatively small compared to the total.
    - Urban males ("M, Urban") also show a similar pattern, with a high count but a modest conversion proportion.
    - Rural females ("F, Rural") and rural males ("M, Rural") have lower total counts but similar conversion proportions.
2. **Conversion Patterns**
    - Across all categories, the conversion rate (indicated by the purple segment) is relatively low compared to non-conversions.
3. **General Trends**:
    - Urban areas have higher counts for both genders, but conversion rates remain low across all groups.
    - Rural areas show lower total counts, with a marginally better conversion proportion

This chart suggests that while urban areas have higher representation, the likelihood of conversion does not differ significantly across gender or area

4. Impact of Income & Age on Conversion



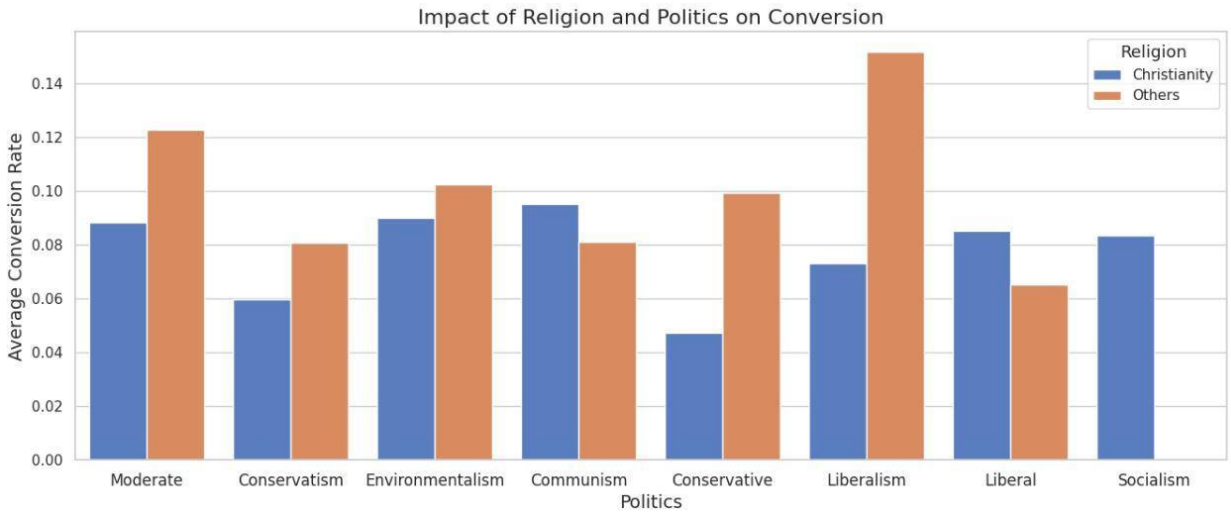
This stacked bar chart shows the count of "True Conversion" (Yes or No) across different income and age categories.

Key observations:

1. **Income under 100K:**
    - In the 55-64 age group, the highest count is observed, with most conversions marked as "No" and a small number as "Yes."
    - The 25-34 and 45-54 age groups also have a notable count, primarily "No" with few conversions marked as "Yes."
    - Younger age groups (18-24) have a significantly lower count compared to the older age groups.
2. **Income over 100K:**
    - Similar to the under 100K group, the 55-64 age group has the highest count, with most conversions marked as "No."
    - Other age groups (25-34 and 45-54) have lower counts, with more "No" responses and very few "Yes" responses.

Across both income categories, the 55-64 age group has the highest count of responses, with most responses marked as "No." There are relatively few conversions marked as "Yes" in any age and income group, suggesting that conversion rates may be low across these segments.

5. **Impact of Religion & Politics on Conversion**



This bar chart illustrates the relationship between political affiliation, religion (Christianity vs. Others), and average conversion rates.

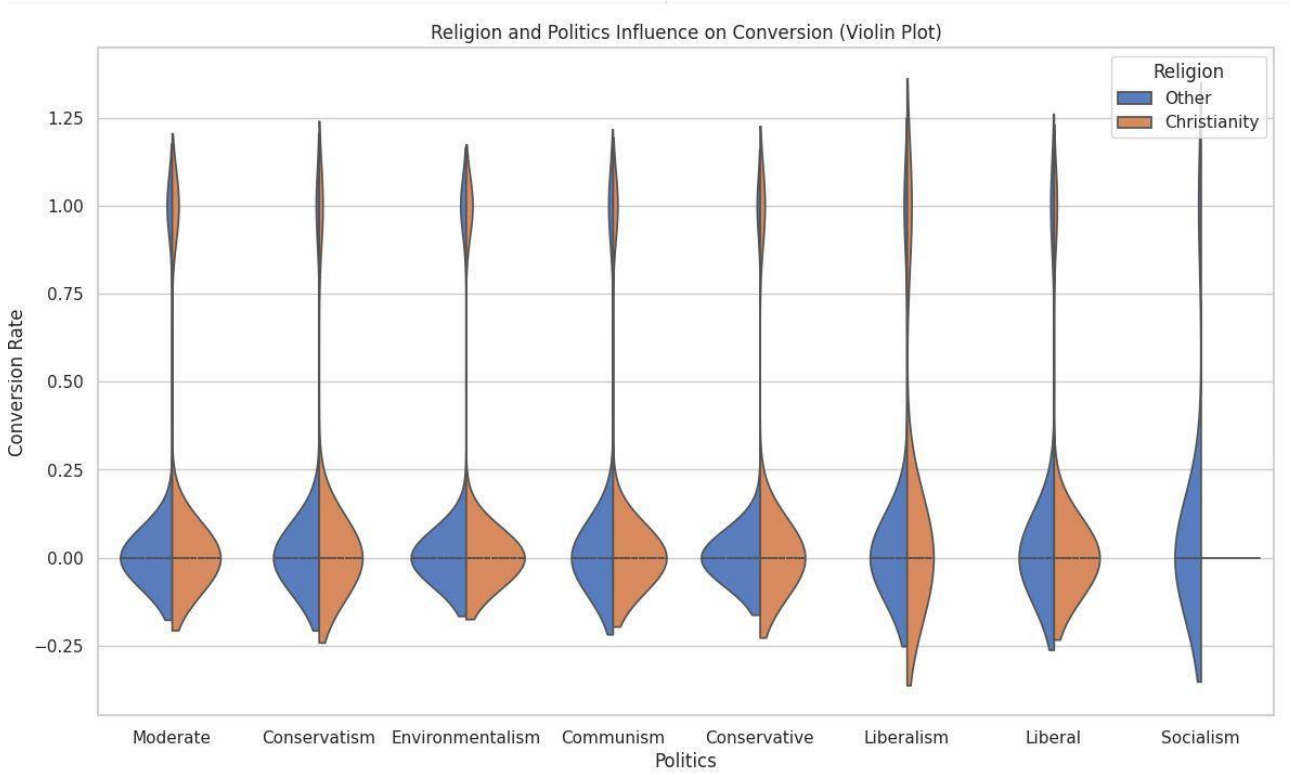
Key observations:

1. **Conversion Rate by Religion:**
   - In most political categories, individuals identifying with "Others" have a higher average conversion rate than those identifying with "Christianity."
   - The disparity is particularly noticeable in "Liberalism," where the conversion rate for "Others" is significantly higher than that for "Christianity."
2. **Political Influence on Conversion:**
   - Political affiliations such as "Moderate," "Environmentalism," and "Liberalism" show higher conversion rates for "Others" than for Christians.
   - Conversion rates are relatively low for both religious groups among "Conservative" and "Liberal" political affiliations.

3. **Highest Conversion Rates:**
   - The highest conversion rate is seen in "Liberalism" for "Others," followed by "Moderate" and "Communism" for "Others."

Political affiliation and religion both impact conversion rates, with "Others" generally showing higher average conversion rates than "Christianity" across various political categories, especially in "Liberalism." This suggests that individuals with certain religious and political affiliations may respond differently to conversion opportunities.

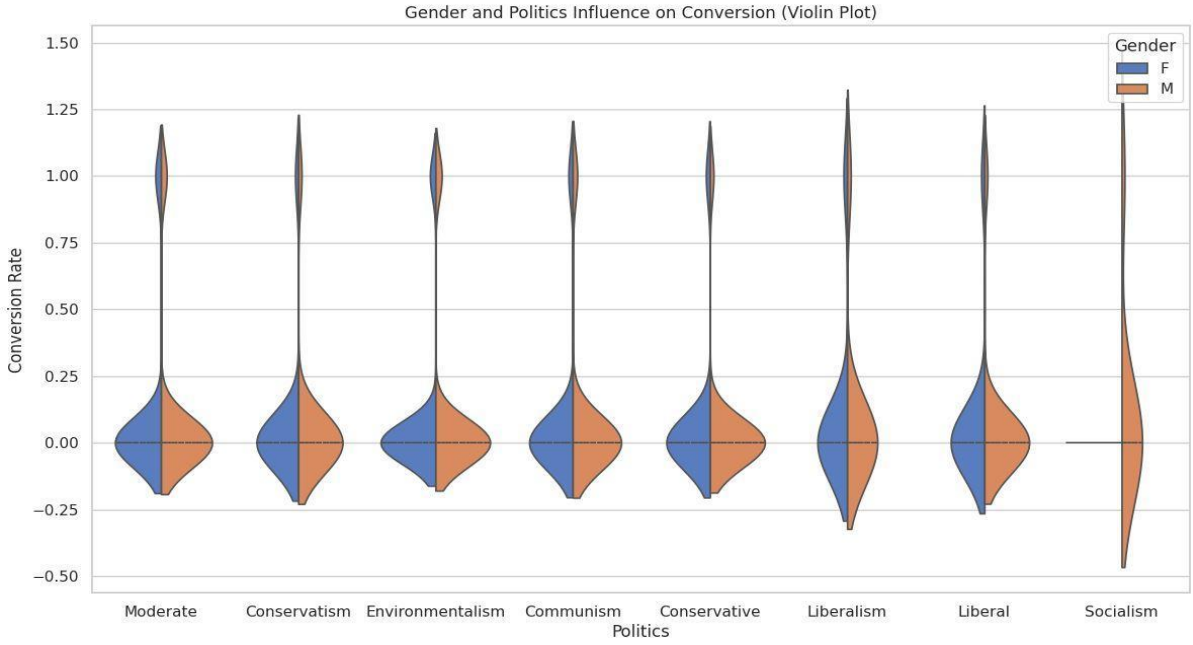
6. Violin Plot indicating Conversion based on Religion and Politics



Key observations:

- This plot shows the conversion rates across various political affiliations, segmented by religion (Christianity vs. Other religions).
- Christianity appears to have a slightly wider distribution across conversion rates in some political categories, particularly Environmentalism and Liberalism.
- Conversion rates for both religions remain close to zero across most political categories, indicating that political affiliation might not strongly influence conversion, though there is some variation.

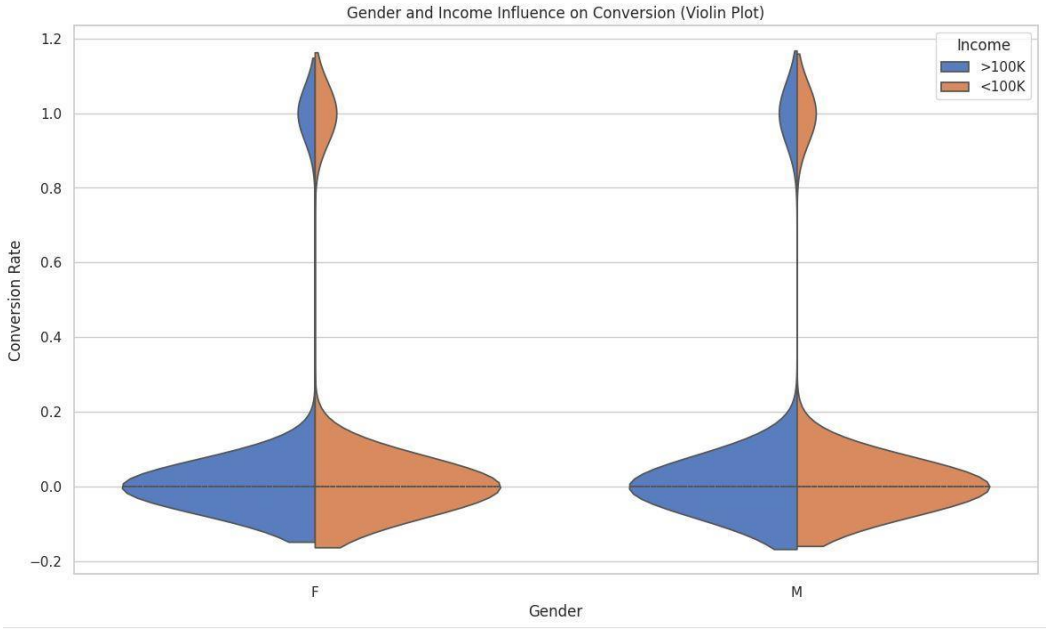
7. Violin Plot indicating Conversion based on Gender and Politics



Key observations:

- This plot compares conversion rates for males and females across political beliefs.
- The distributions are relatively similar between genders across most political categories.
- The most noticeable variation occurs in the Socialism category, where males show a broader distribution with some conversion rates reaching higher levels compared to females.
- Generally, the conversion rate remains around zero, suggesting a weak influence of gender within each political category on conversion rates.

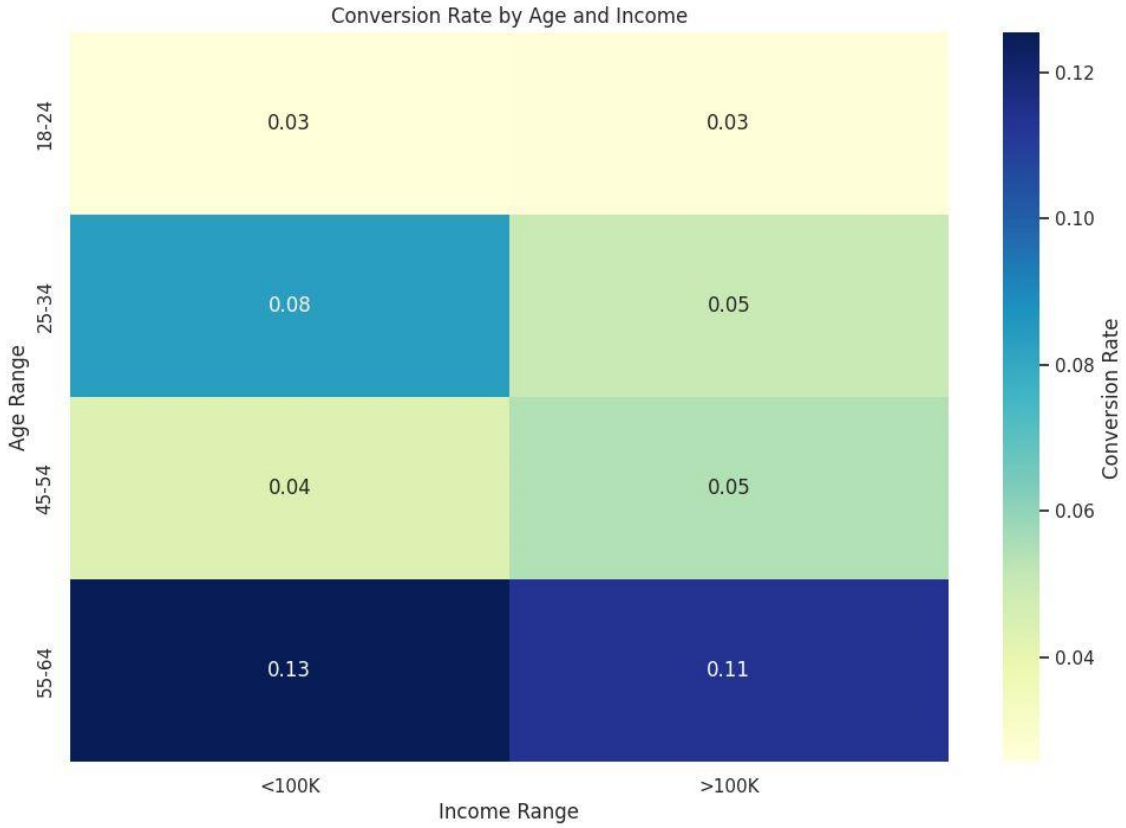
8. Violin Plot indicating Conversion based on Gender and Income



Key observations:

- This plot examines the impact of income (above or below $100K) and gender on conversion rates.
- There is a higher concentration of conversion rates around zero, especially for those earning below $100K.
- Individuals earning above $100K have slightly more variation, with both genders showing similar distributions, though with some differences in higher conversion rates.
- This suggests income might have a minor influence on conversion, with high-income individuals showing more variance in conversion rates compared to lower-income individuals.

9. Comparison of Conversion Rates (True and Predicted) based on several parameters



The heatmap displays the conversion rates across different age and income groups.

Key observations:

Age:

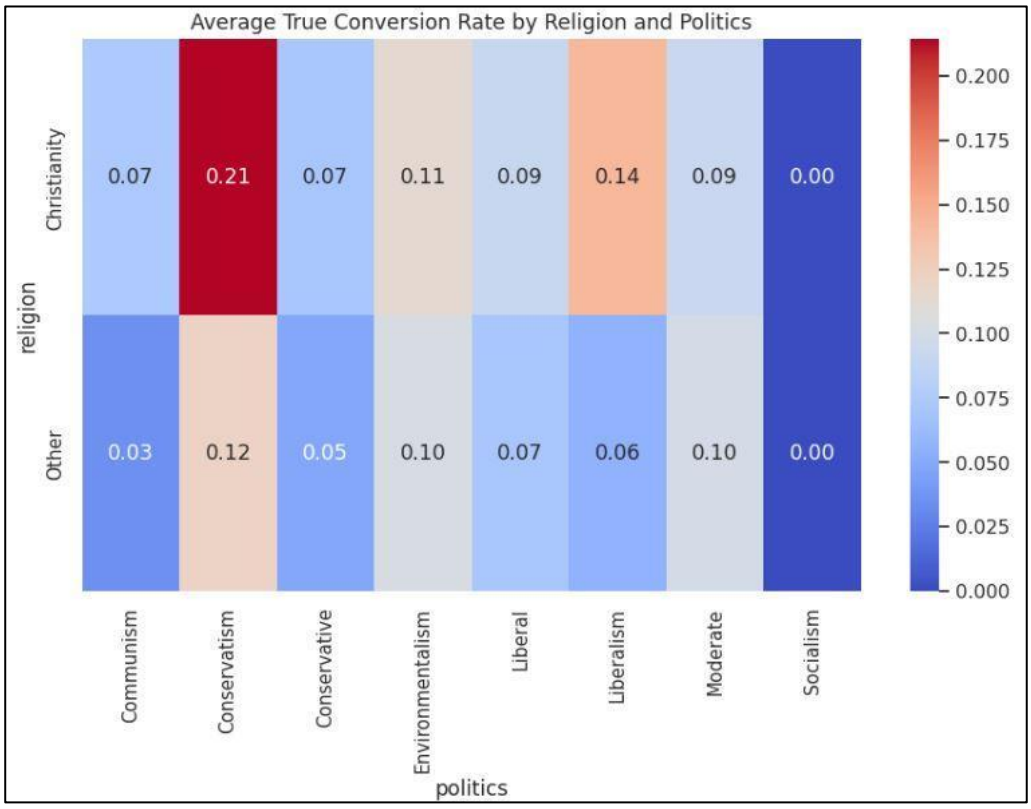
- The conversion rate seems to increase with age. The highest conversion rate is observed in the 55-64 age group, followed by 45-54 and 25-34. The 18-24 age group has the lowest conversion rate.

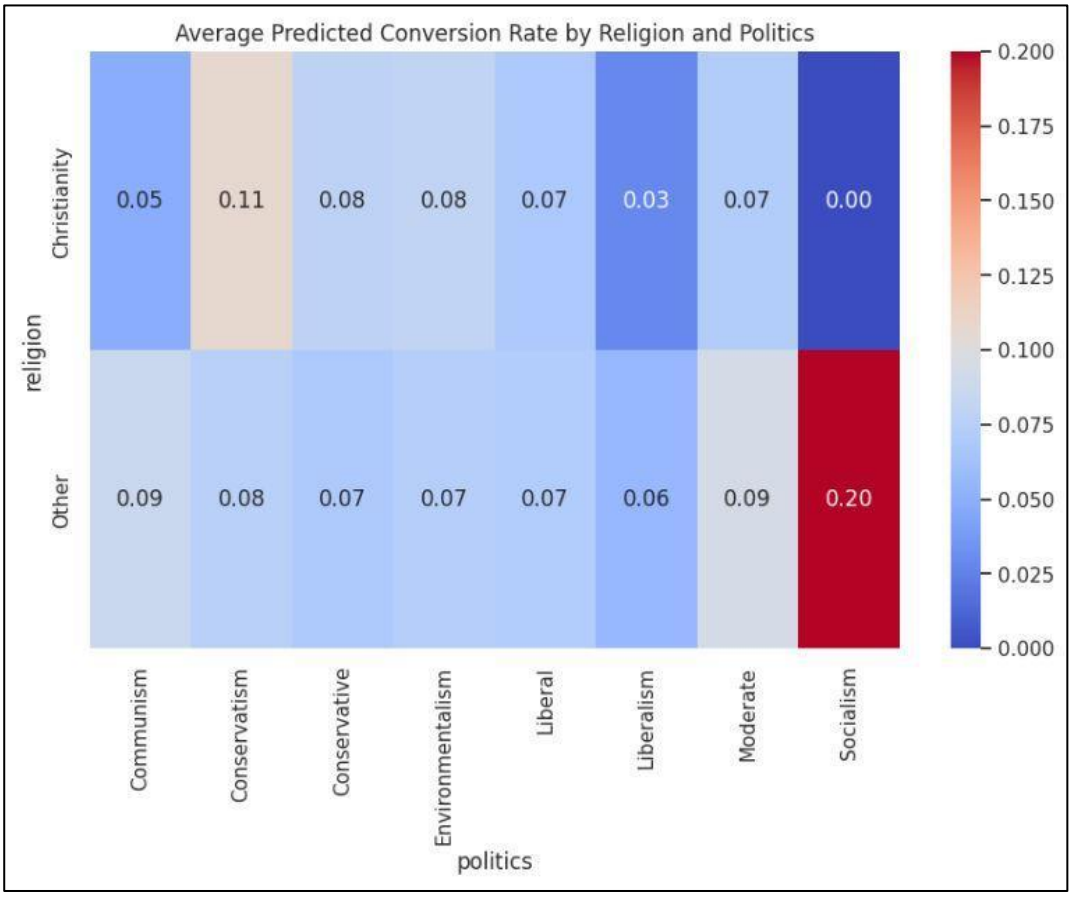
Income:

- The conversion rate is generally higher for the income group >100K compared to the <100K group.

Overall:

- The highest conversion rate is seen in the 55-64 age group with an income >100K.
- The lowest conversion rate is seen in the 18-24 age group with an income <100K.



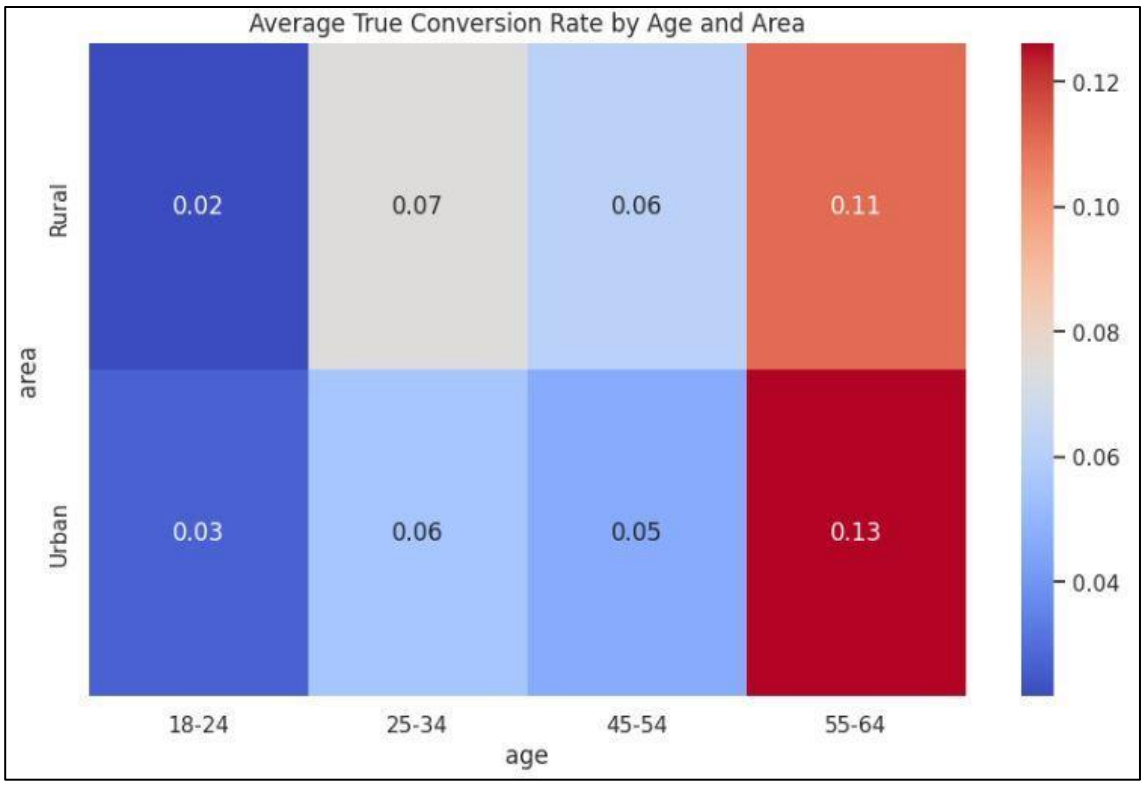


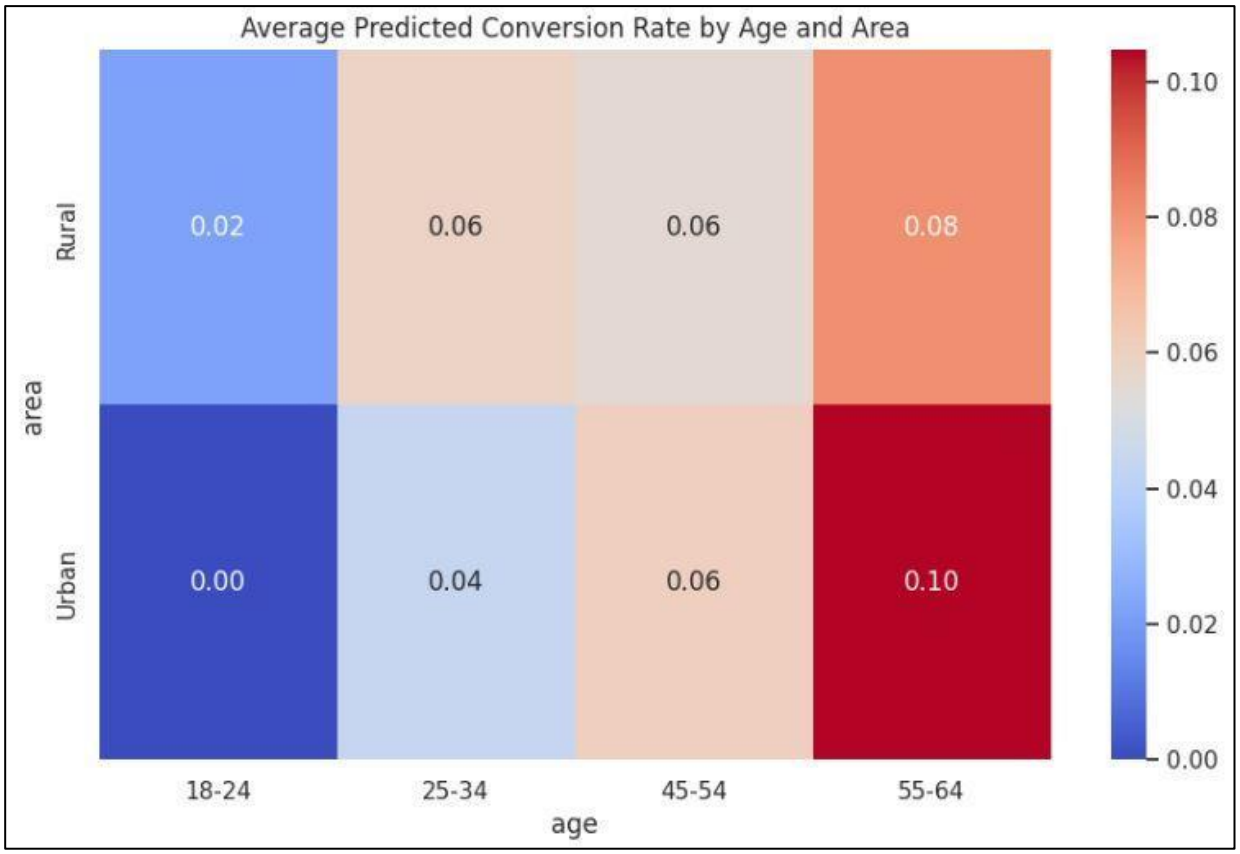
Key observations:

- **Political Affiliation Impact:** The heatmap suggests that political affiliation plays a significant role in conversion rates. Individuals with "Conservative" and "Moderate" political views generally show higher conversion rates compared to those with "Liberal" views, regardless of their religious affiliation.

- **Religion Impact:** While religion seems to have less of an overall impact than political affiliation, there are some notable differences. For example, Christians, particularly those with "Conservative" or "Moderate" political views, tend to have higher conversion rates compared to those with "Other" religious beliefs.

- **Combined Effect:** The most significant conversion rates appear in the cells where Conservative or Moderate politics intersect with Christianity. These groups seem most likely to convert based on the data.





The heatmap visualises the average conversion rates for different age groups and areas (rural/urban).

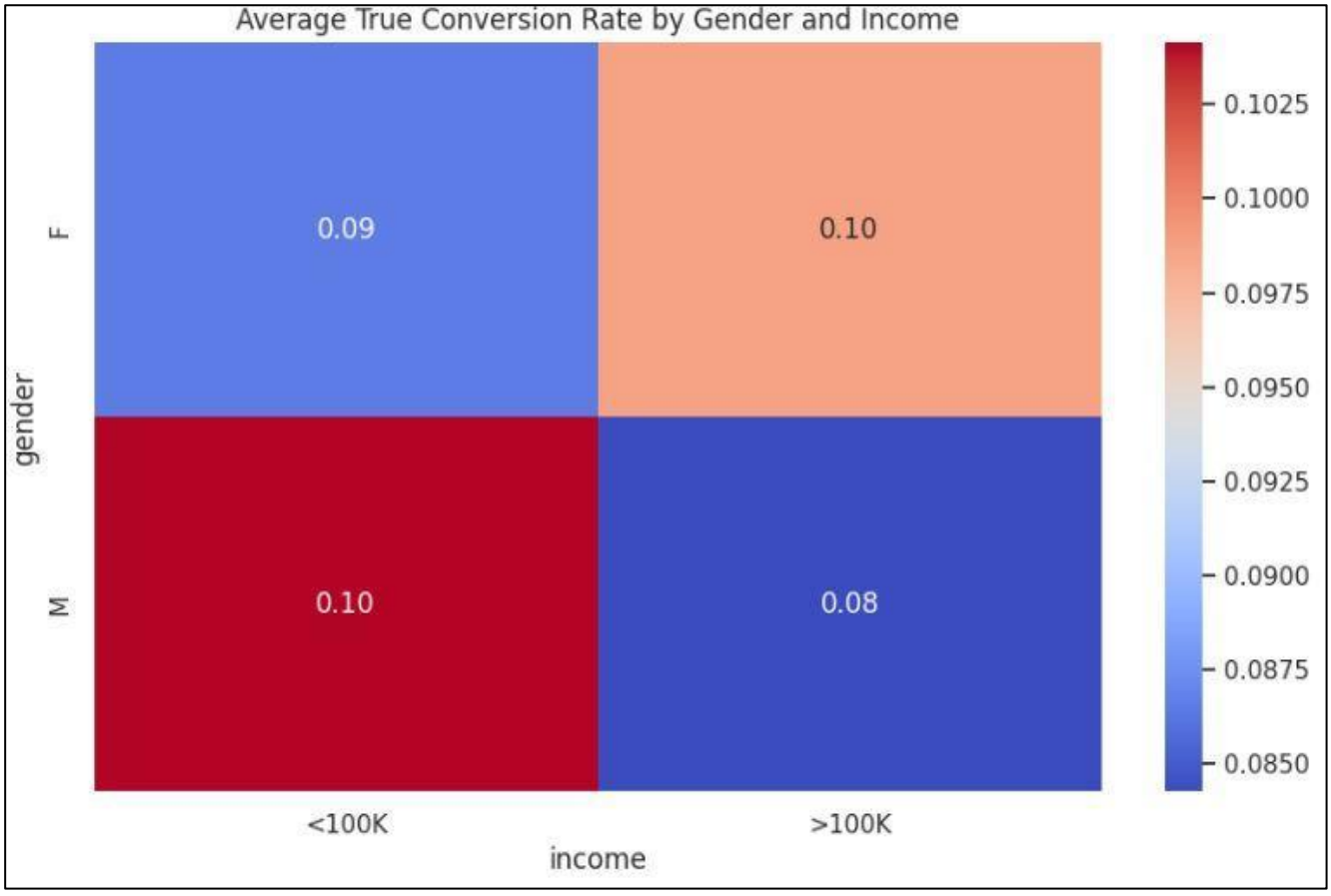
Key observations:

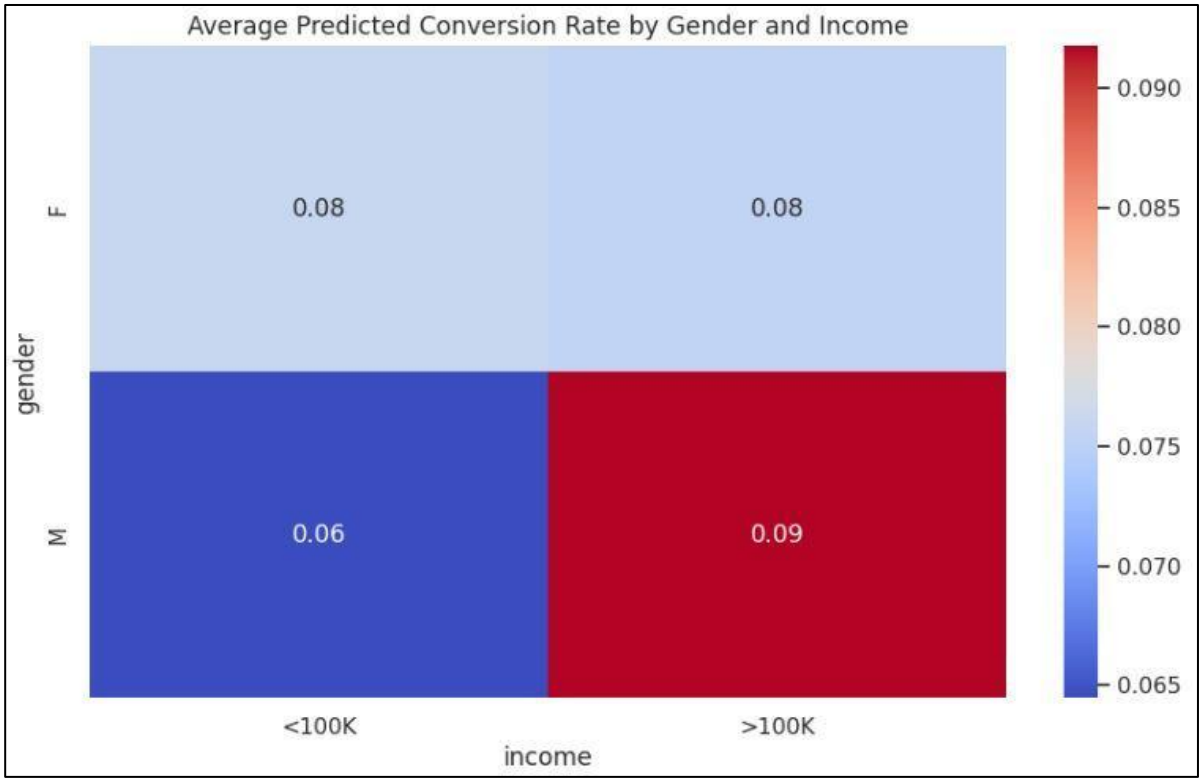
1. **Higher Conversion in Urban Areas:** Generally, urban areas tend to have higher conversion rates compared to rural areas across most age groups. This could indicate that the target audience might be concentrated in urban areas.

2. **Age-Based Variations:**
   - 35-44 Age Group: This age group shows the highest conversion rates, especially in urban areas. This might be the key demographic for the ad campaign.
   - Younger Age Groups (18-24, 25-34): They have relatively lower conversion rates in both rural and urban areas, but slightly better in urban.
   - Older Age Groups (45-54, 55-64, 65+): Their conversion rates are generally lower. It's worth considering if the ad is resonating with older demographics.

3. **Specific Observations:**
   - In the heatmap for true conversion, there's a noticeable spike in conversion for the 35-44 age group in the urban area, suggesting a potential sweet spot for ad targeting.

- The pattern for predicted conversion is similar but slightly less pronounced. The model might not be perfectly capturing the differences between age groups and areas.

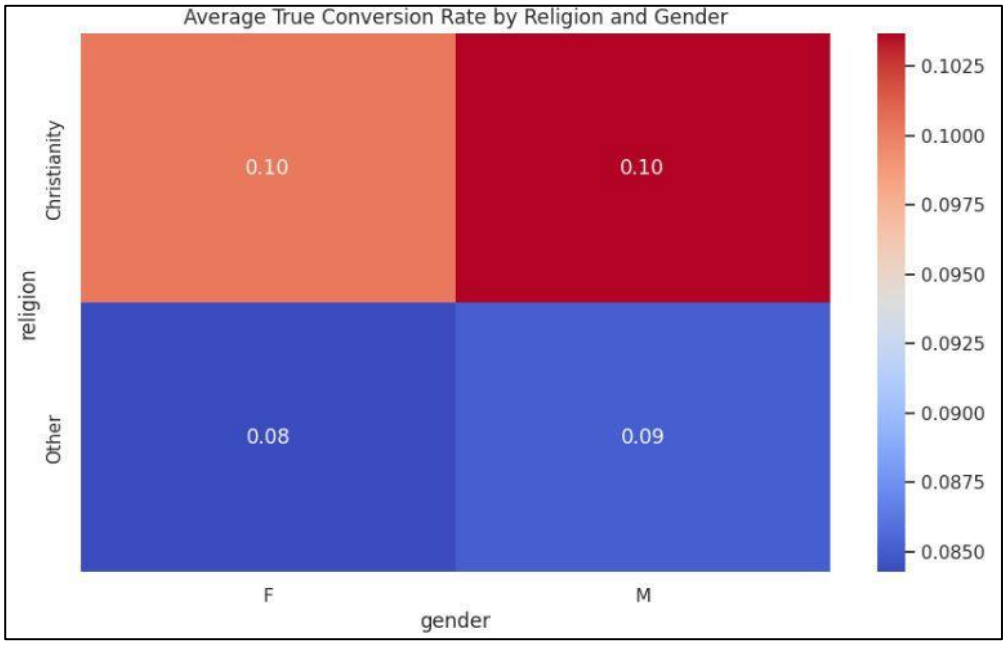


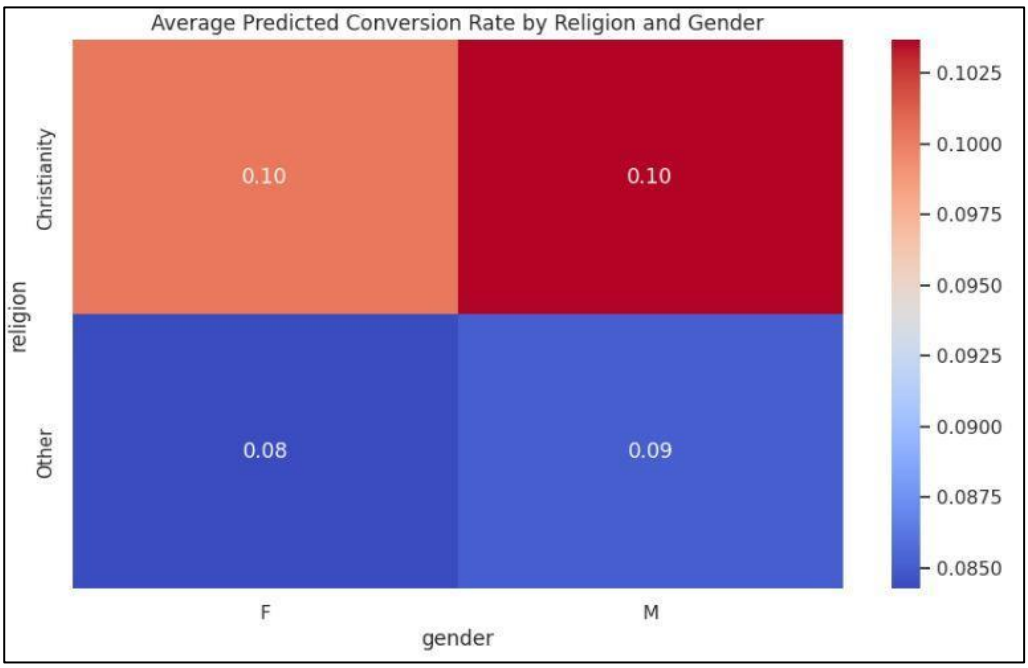


Key observations:

1. **Income is a strong predictor of conversion:** Higher income levels consistently show higher conversion rates for both genders, indicating a positive correlation between income and the likelihood of conversion. This observation holds true for both actual and predicted conversion rates.

2. **Potential gender differences within income groups:** While income is the primary driver, subtle differences in conversion rates might exist between genders within specific income brackets. These variations are worth exploring further to understand potential gender-specific targeting opportunities.





1. Religion and gender may have an interaction effect on conversion: The heatmap reveals potential differences in conversion rates based on the specific combination of religion and gender. This suggests that these two factors might interact to influence the likelihood of conversion, rather than acting independently.

2. Christian females may exhibit a higher conversion tendency: Based on the heatmap, Christian females might have a slightly higher average conversion rate compared to other religious and gender groups. However, the magnitude of this difference might not be substantial, and further analysis is needed to confirm its statistical significance.

3. Data Mining Algorithms and Result Analysis

The Summary of Data Mining Algorithms used as part of the Project are:

- Decision Tree Model (Pre-Pruned & Post-Pruned)
- Bagging
- Random Forest
- Naive Bayes Classifier
- Artificial Neural Network
- XGBoost

The metrics used to evaluate the performance of these classification models which are also indicated as part of the report are

- Accuracy of the Model
- Receiver Operating Characteristic Curve
- Confusion Matrix
    - True Positive Rate / Sensitivity / Recall
    - True Negative / Specificity
    - False Positive
    - False Negative
    - Precision
    - F-Score

In our scenario, the inference of each of the 4 terms in the confusion matrix are provided below and the interpretation is the same across the cases.

- True Positive - Predicting correctly that a customer will be converted
- True Negative - Predicting correctly that a customer will not be converted
- False Positive - Predicting incorrectly that a customer will be converted
- False Negative - Predicting incorrectly that a customer will not be converted

From the perspective of the importance, the most important metrics to assess the model performance shall be

- Lower values of False Negative Rate (to not lose out on potential customers)
- High values of True Positive Rate (if required can give strategic discounts or design products as appropriate)

- Several strategies can be thought of and built around each group of customers to enhance the conversion rate.

Decision Tree Model and Result Analysis:

1. Data Preparation

- **Target and Features Selection**: The *true_conversion* column is defined as the target variable, representing whether a customer converts or not. The rest of the dataset, excluding the *predicted_conversion* column, serves as the features.

- **One-Hot Encoding of Categorical Data**: To make categorical features compatible with the model, one-hot encoding is applied. This converts categorical variables into multiple binary columns, facilitating a more effective decision tree model by preventing the misinterpretation of ordinal relationships.

2. Splitting the Data

- **Train-Test Split:** The dataset is split into training and testing sets, with 70% of the data allocated for training and 30% for testing. This split allows the model to train on a substantial portion of the data while reserving unseen data for evaluation, providing an unbiased assessment of the model's predictive performance.

3. Initial Decision Tree Model Training

- **Model Training without Depth Constraints:** The first decision tree model is trained without specifying a maximum depth, allowing it to grow fully. This approach aims to understand the decision boundaries formed and to compare with subsequent models that incorporate depth constraints.

- **Tree Visualization:** The full decision tree is visualised, providing insight into the feature splits and depth of the tree. This visualisation aids in understanding the complexity of the model and how it partitions the feature space based on different attributes.

4. Initial Model Evaluation

- **Predictions and Probability Scores:** The trained model is applied to the test set, generating both predicted labels and probability scores. These predictions allow for the calculation of various evaluation metrics.

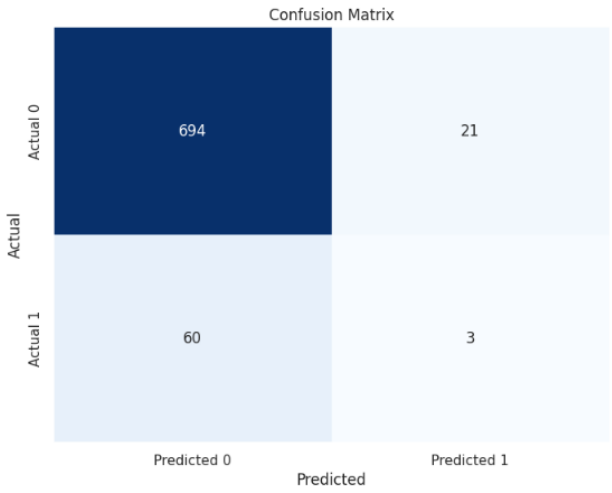
- **Confusion Matrix and Heatmap:** A confusion matrix is created to assess the number of true positives, true negatives, false positives, and false negatives. A heatmap visually represents these metrics, facilitating a quick understanding of the model's performance.

- **ROC Curve and AUC:** The Receiver Operating Characteristic (ROC) curve is plotted, illustrating the trade-off between the true positive rate and the false positive rate. The Area Under the Curve (AUC) quantifies the model's performance, with a higher AUC indicating better discrimination between positive and negative classes.

5. Hyperparameter Tuning - Finding Optimal Depth

- **Cross-Validation for Depth Selection**: The maximum depth of the tree is optimised by evaluating the AUC score across a range of depths using cross-validation. The depth that yields the highest mean AUC score is chosen as the optimal depth, balancing model complexity and predictive power.

- **Final Model Training with Optimal Depth**: A decision tree with the best-performing depth is restrained on the full training set. This tuned model is expected to generalise better on unseen data, mitigating overfitting associated with deeper trees.



Inference:

The area under the ROC curve is equal to 0.55 which is very similar to the result of Random guessing. So some improvement measures need to be performed.

6. Pruning the Decision Tree (CCP Alpha Tuning)

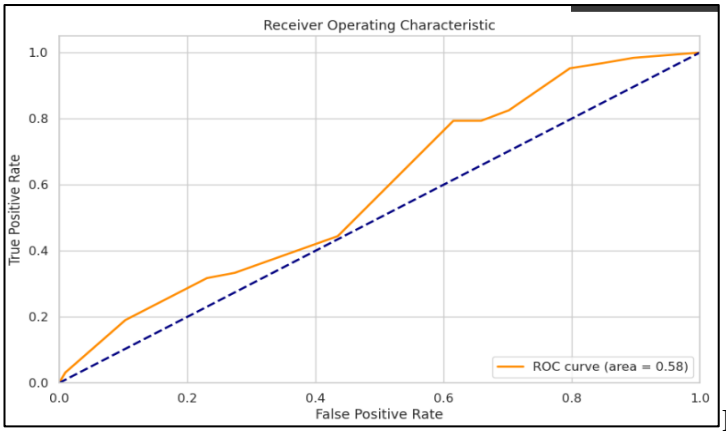
- **Cost Complexity Pruning**: The decision tree is further optimised using Cost Complexity Pruning (CCP) by tuning the *ccp_alpha* parameter. Different values of ccp_alpha yield a series of pruned trees, each reducing the complexity of the tree structure.

- **Validation Accuracy and Pruning Selection**: Each pruned tree's accuracy is assessed on the validation set. The *ccp_alpha* value with the highest validation accuracy is selected, and the final pruned tree is retrained. This approach ensures that the tree remains interpretable without sacrificing too much accuracy.

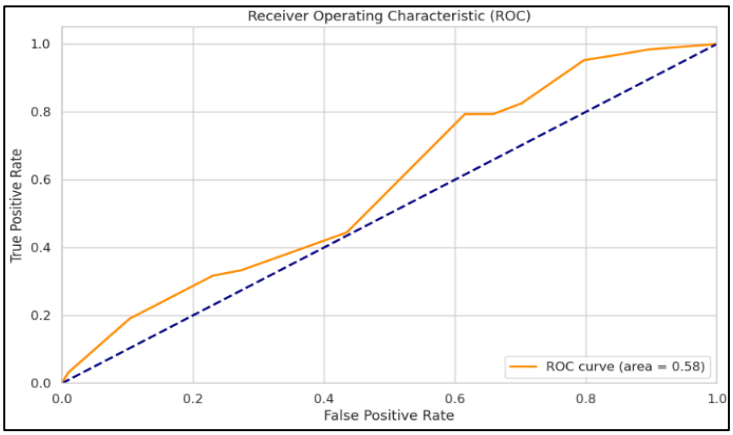
7. Evaluation of the Pruned Decision Tree

- **Test Set Accuracy and Confusion Matrix:** The final pruned tree model is evaluated on the test set, and its accuracy is reported. A confusion matrix and heatmap are generated to compare actual and predicted labels post-pruning, highlighting the model's classification accuracy.

- **ROC Curve and AUC for Pruned Tree:** The ROC curve for the pruned model is plotted, with the AUC score calculated to compare its performance against the initial and optimised decision tree models.



Pre pruned



Post pruned

Inference:

The ROC- AUC came out to be 0.58 which is not much improvement in performance of the model compared to simple decision tree model.

Bagging Classifier Model and Evaluation

1. Model Setup and Hyperparameter Tuning

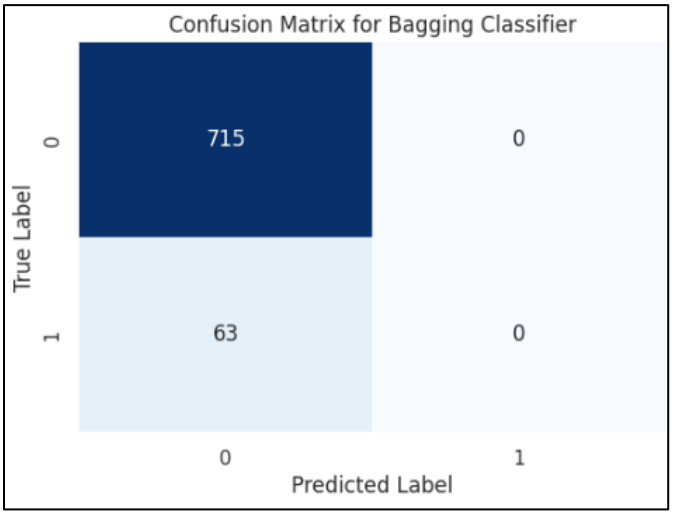
- **Initialization**: A Bagging Classifier is instantiated with *random_state=42* to ensure reproducibility.
- **Parameter Grid**: To optimise performance, the model undergoes grid search using a comprehensive parameter grid:
    - *n_estimators*: Number of base estimators (50, 100, 200).
    - *max_samples*: Proportion of samples used for training each estimator (0.6, 0.8, 1.0).
    - *max_features:* Proportion of features used for training each estimator (0.5, 0.75, 1.0).
    - *bootstrap:* Determines whether sampling is done with replacement.
- **Grid Search Cross-Validation**: Conducted with 5-fold cross-validation using *roc_auc* as the scoring metric. The best parameter configuration is identified based on AUC.

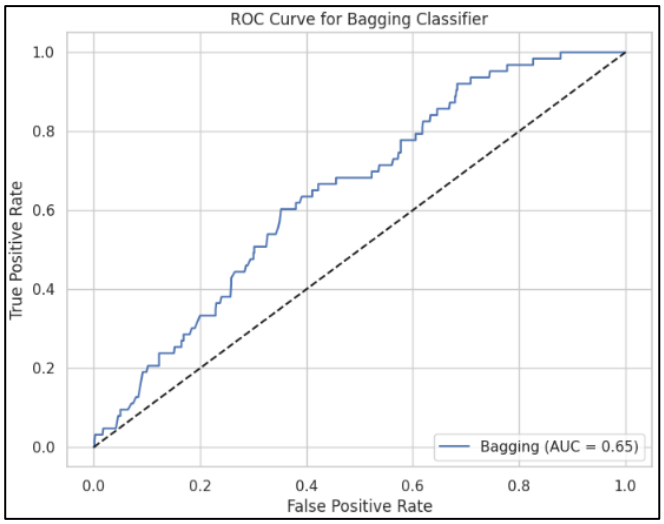
2. Model Training and Prediction

- **Best Model Selection**: The optimised Bagging Classifier, selected from grid search, is then trained on the training data.
- **Prediction on Test Data**: Predictions and probabilities for the test set are obtained for model evaluation.

3. Model Evaluation Metrics

- **Classification Report and AUC Score**: The classification report (precision, recall, F1-score) and the ROC AUC score are presented for the test predictions, summarising model performance.
- **Confusion Matrix**: A heatmap of the confusion matrix illustrates correct and incorrect predictions across classes, providing insight into the model's accuracy.
- **ROC Curve**: The ROC curve visualises the trade-off between true positive and false positive rates, with the AUC as an overall performance indicator.





Inference:

The ROC-AUC score obtained is 0.65, which is a good improvement in the performance of the model.

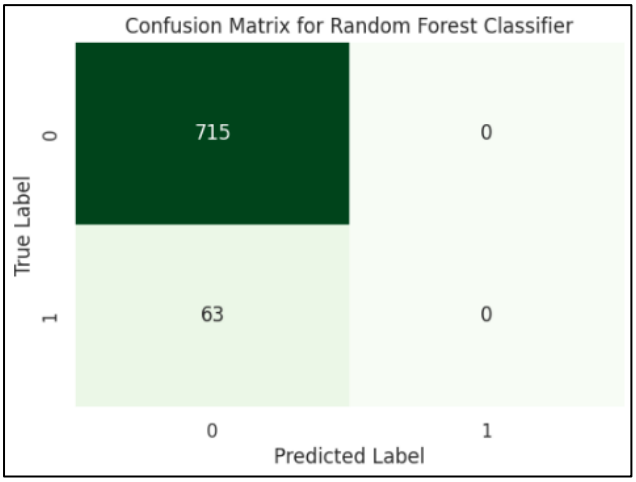
Random Forest Classifier Analysis

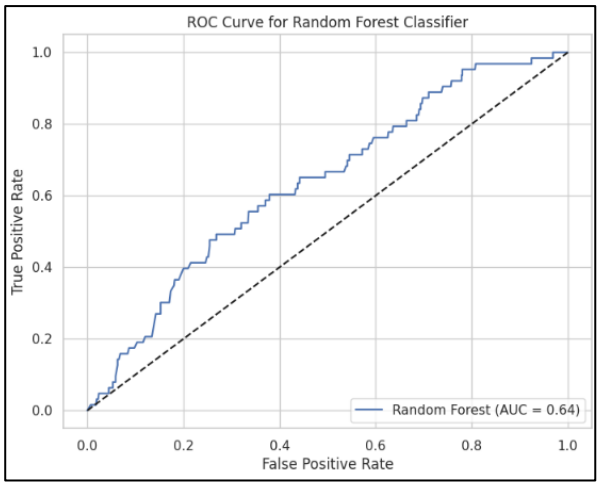
1. Model Setup and Hyperparameter Tuning:

- Initialized a Random Forest Classifier with a comprehensive grid search on parameters such as *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, and *max_features*.
- Used 5-fold cross-validation for *roc_auc* scoring to select the optimal model.

2. Model Evaluation:

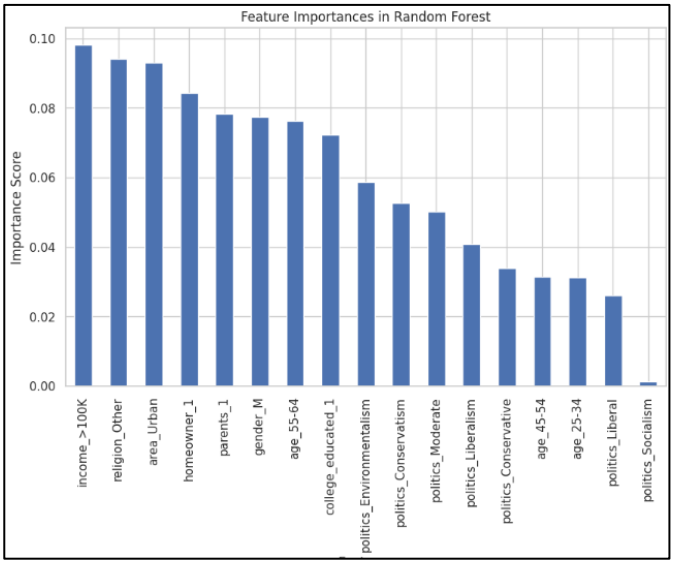
- **Best Parameters and Metrics**: The best configuration and the classification report metrics, including precision, recall, and F1-score, were displayed.
- **Confusion Matrix:** Presented as a heatmap for a quick view of model performance.
- **ROC Curve:** Plotted to evaluate the trade-off between sensitivity and specificity, with an AUC score indicated.





3. Feature Importance:

- The top contributing features were plotted, highlighting key predictors for interpretability.



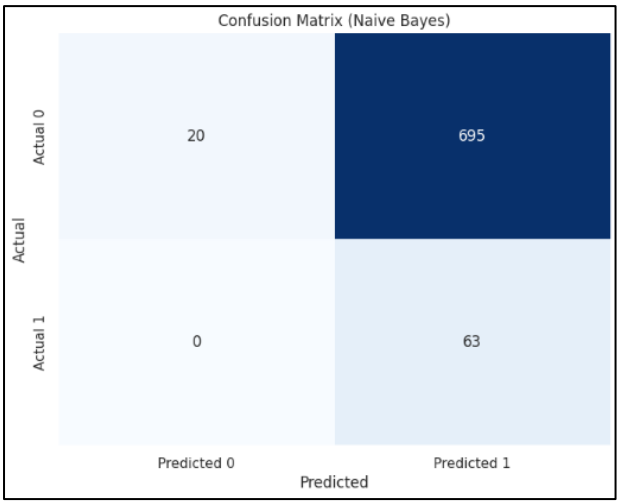
Gaussian Naive Bayes Analysis

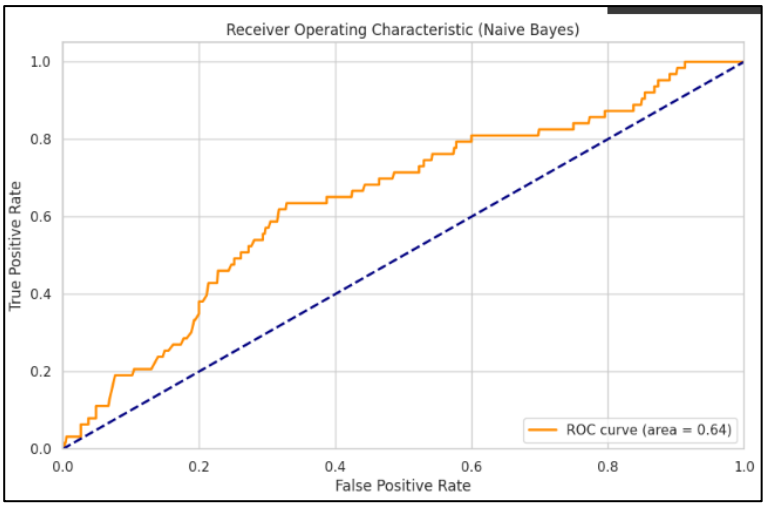
1. Model Training and Prediction:

- Trained a Gaussian Naive Bayes classifier and made predictions on the test set.

2. Model Evaluation:

- **Metrics**: Reported accuracy, precision, recall, F1-score, and ROC AUC to summarize performance.
- **Confusion Matrix**: Visualized to show prediction breakdown between classes.
- **ROC Curve**: Plotted to assess sensitivity vs. specificity, with AUC displayed for an overall view of model effectiveness.





Artificial Neural Network (ANN) Analysis

1. Data Preprocessing and Pipeline:

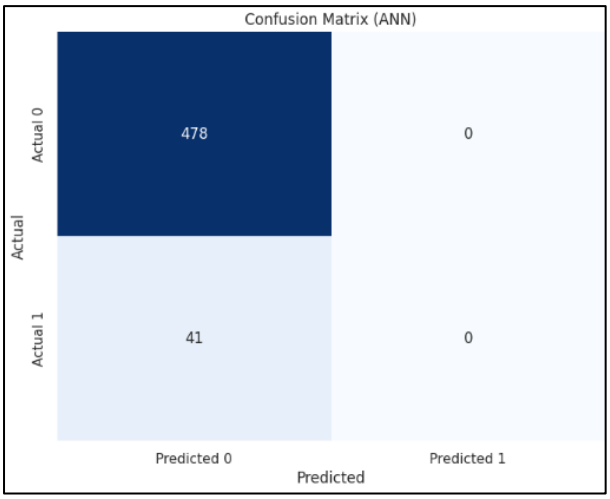
- Prepared data with separate transformations for numerical and categorical features.
- Created a pipeline to standardize features and handle categorical encoding.

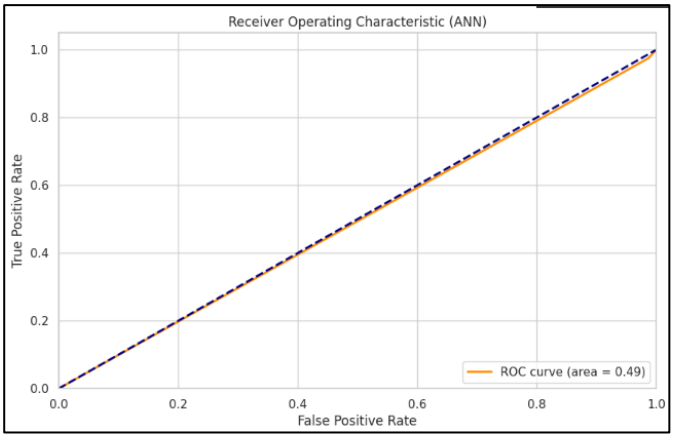
2. Model Architecture:

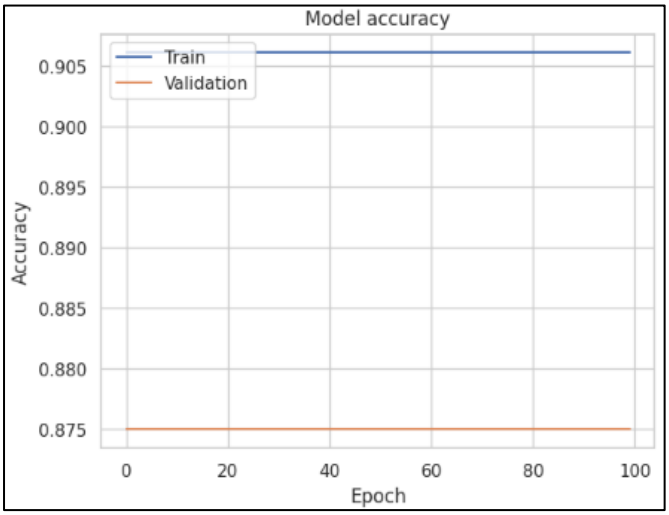
- Designed an ANN with two hidden layers (128 and 64 neurons) and dropout regularization.
- The model was compiled with binary cross-entropy loss and trained for 100 epochs.

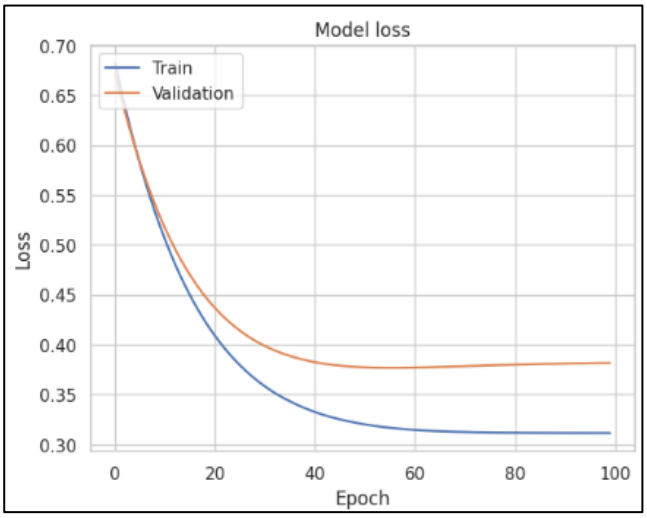
3. Evaluation:

- **Confusion Matrix**: Shows model performance on predictions.
- **ROC Curve**: Visualized to assess model's discriminatory ability (AUC score displayed).
- **Accuracy and Loss Trends**: Training and validation accuracy/loss displayed to monitor model performance over epochs.









The performance of the ANN model on the dataset is ~0.49 which indicates the model performs worse than a random guess.

XGBoost Model Analysis

1. Data Preparation:

- Features converted using one-hot encoding; target label encoded for binary classification.

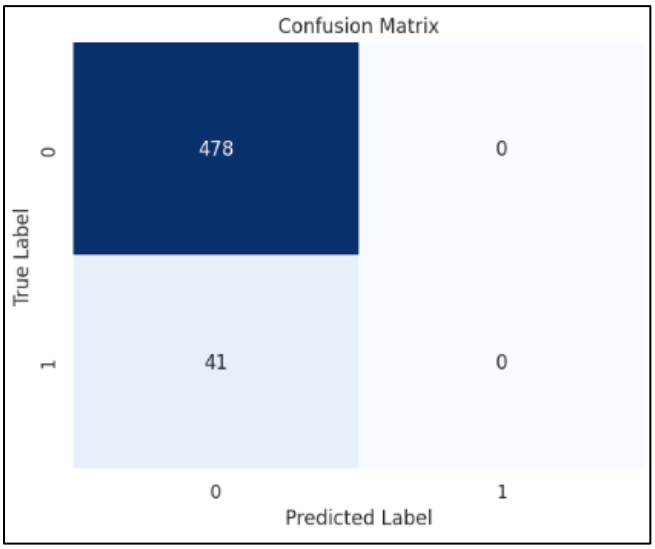
2. Model Tuning and Training:

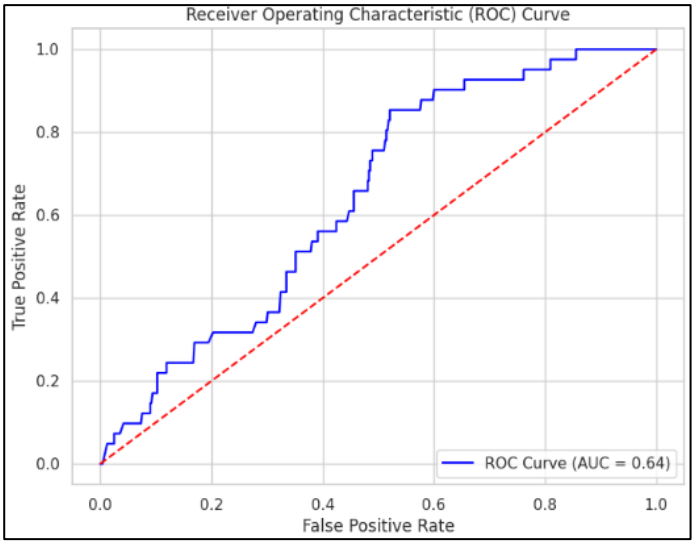
- Used *RandomizedSearchCV* to optimize hyperparameters with 50 iterations and 5-fold cross-validation.

3. Evaluation Metrics:

- **Confusion Matrix**: Assesses prediction accuracy between true and predicted classes.
- **ROC Curve**: Shows model performance; AUC value indicates discriminative ability.

- **Classification Report**: Provides precision, recall, and F1 scores.





Results

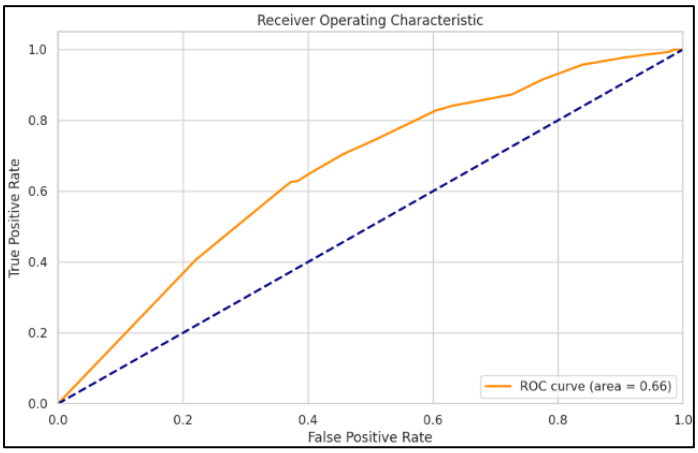
- **Best Parameters**: Displayed post-RandomizedSearchCV.
- **Performance Scores**: Precision, recall, F1-score, and AUC provide insights into model effectiveness.

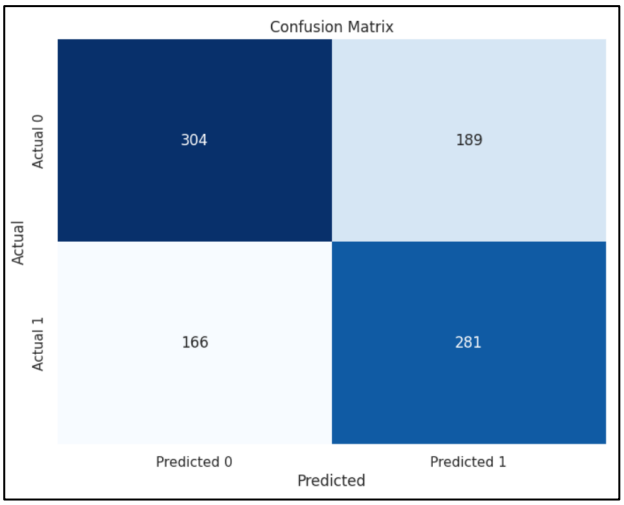
	precision	recall	f1-score	support
0	0.92	1.00	0.96	478
1	0.00	0.00	0.00	41
accuracy			0.92	519
macro avg	0.46	0.50	0.48	519
weighted avg	0.85	0.92	0.88	519

The performance of the models is not very high or satisfactory and the reason can be due to the class imbalance in the target variable of the “true_conversion”. To overcome this shortcoming, and to improve the performance of the models, dataset is balanced with equal number of true_conversion (0) and true_conversion (1) through the SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic) algorithms. Post balancing the dataset, the performance metrics of each of the models implemented till now are again observed and the performance is compared.

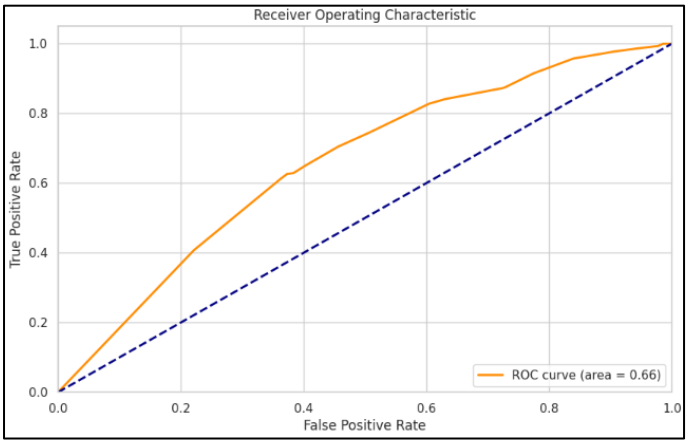
Post SMOTE Results:

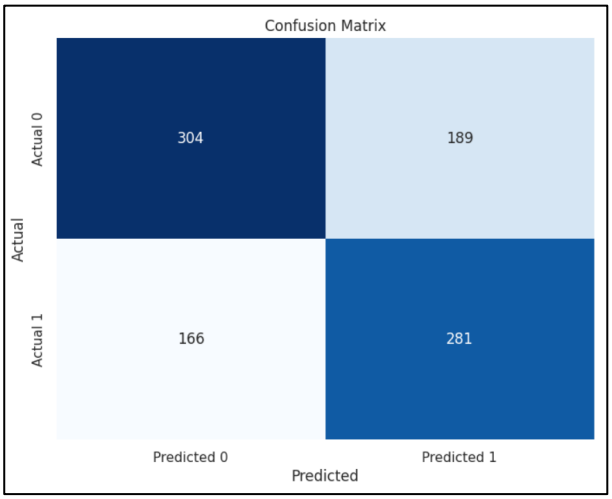
Decision Tree:



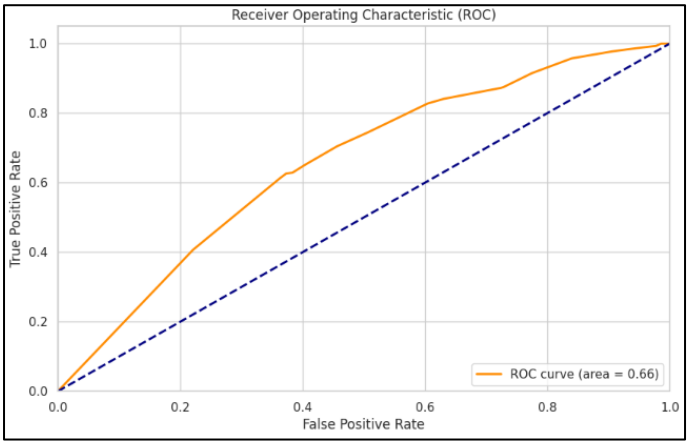


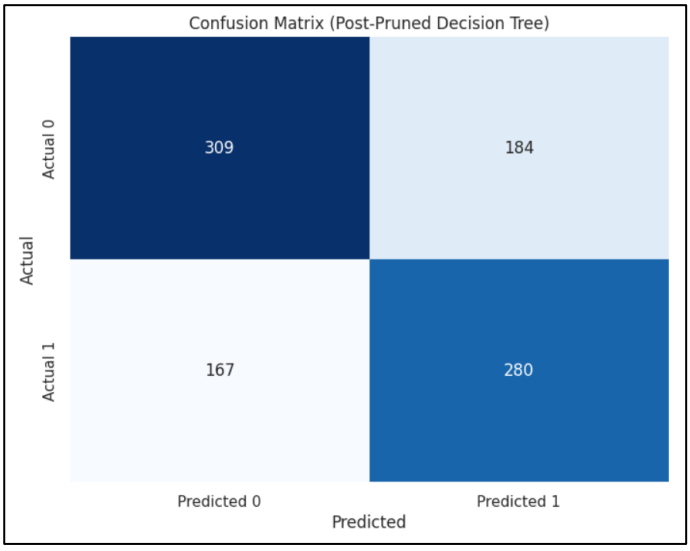
Pre-pruned Decision Tree:



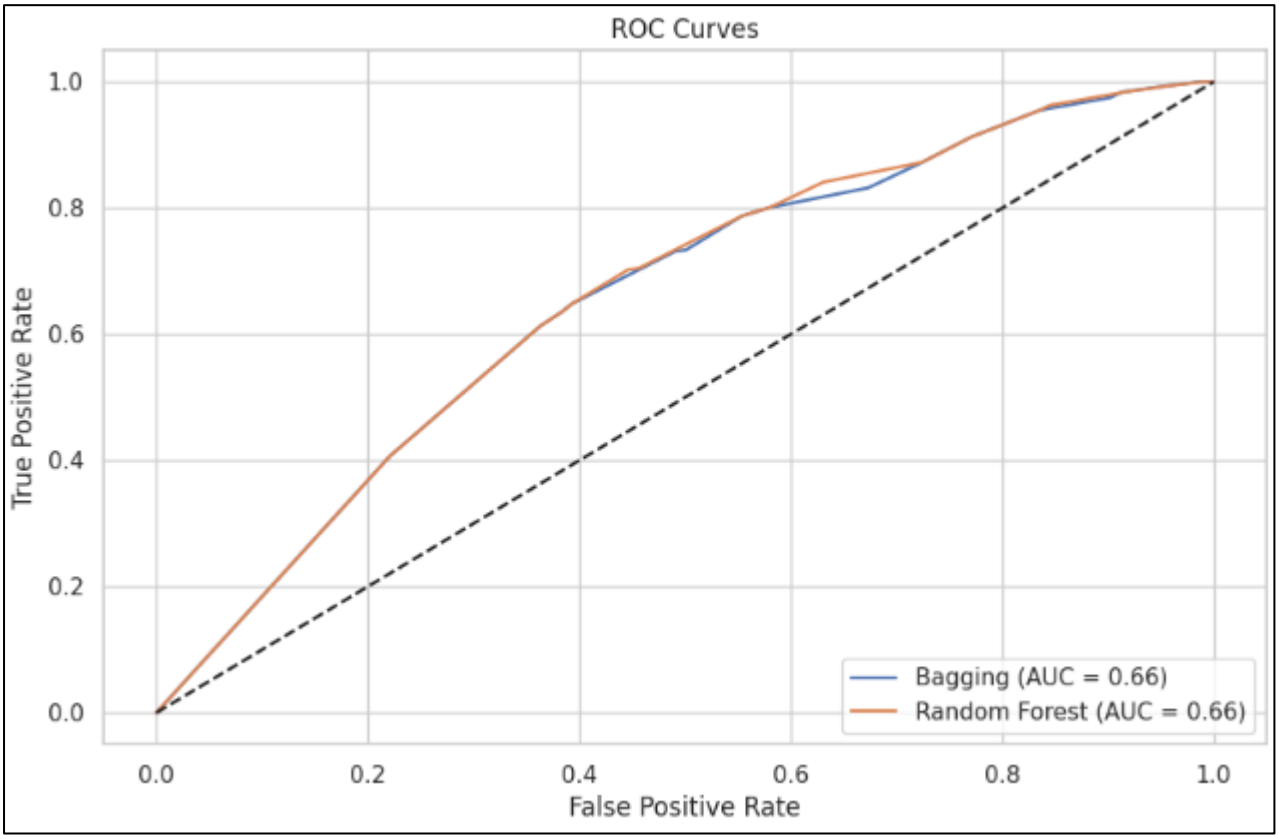


Post pruned Decision Tree:

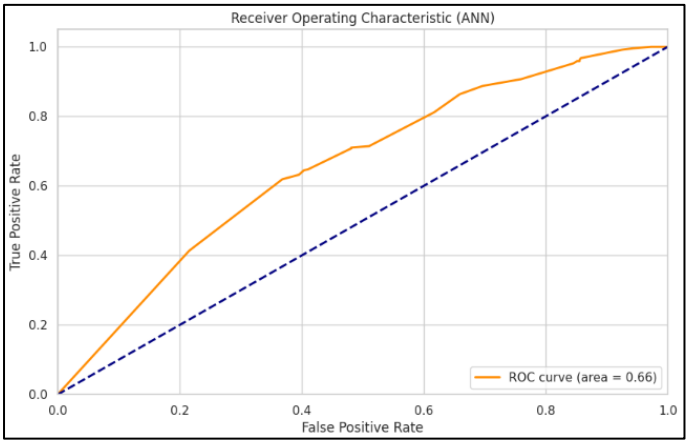


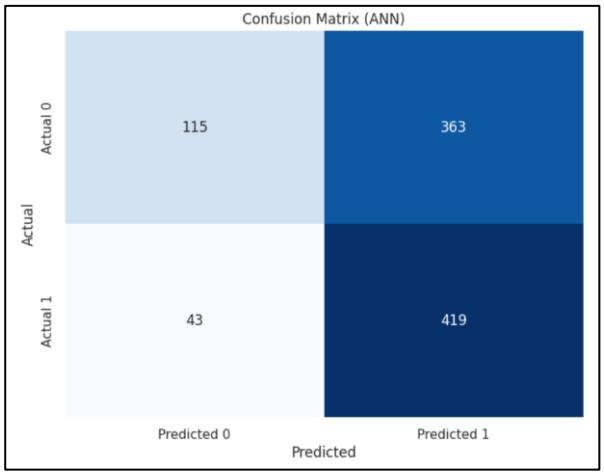


Bagging and Random Forest Classifier:

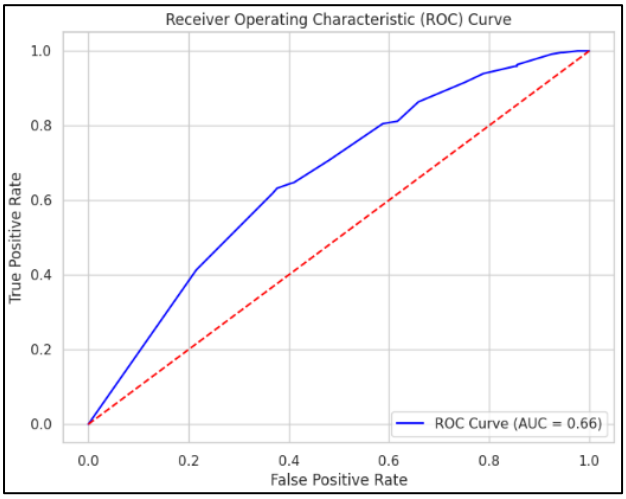


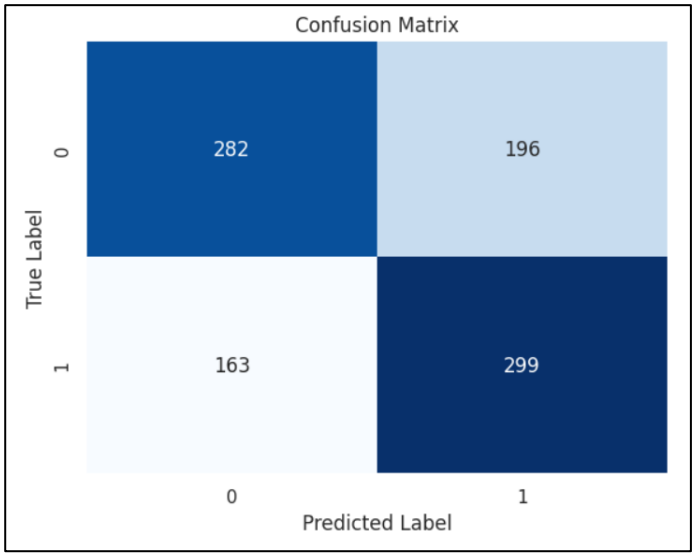
ANN:





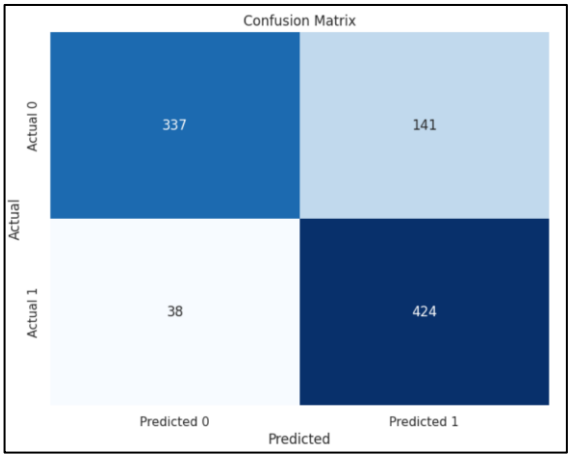
XG Boost:



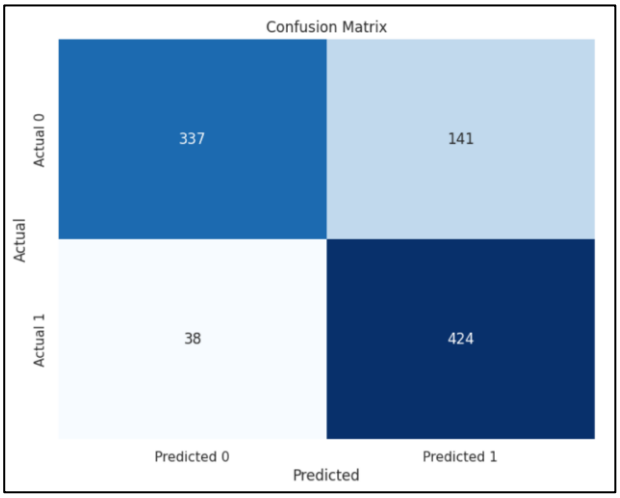


Post ADASYN results:

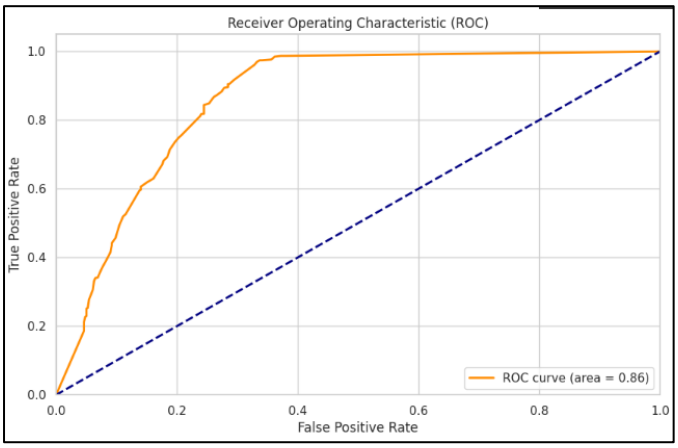
Decision Tree:

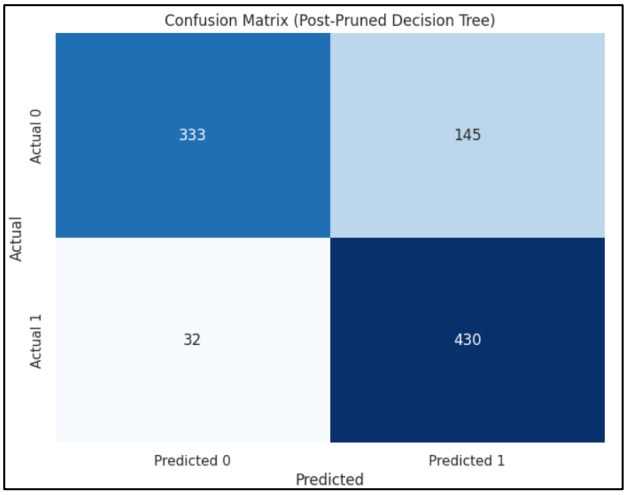


Pre pruned Decision Tree:

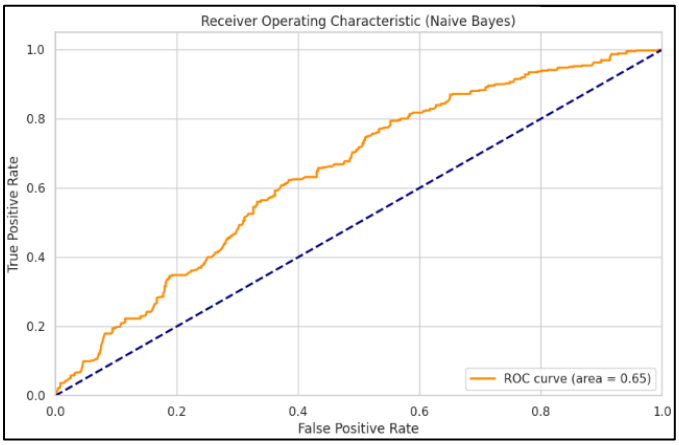


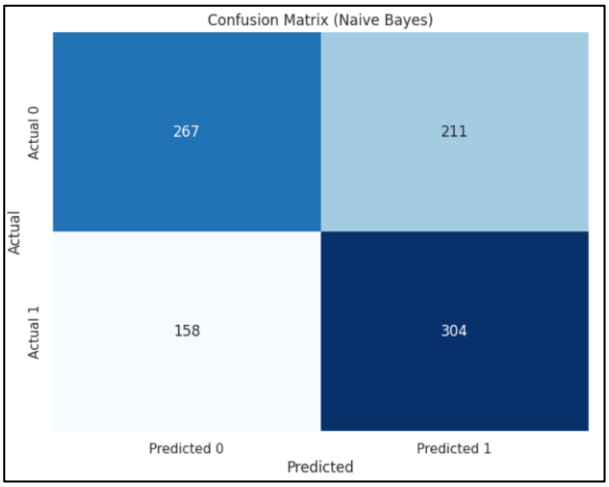
Post pruned Decision Tree:



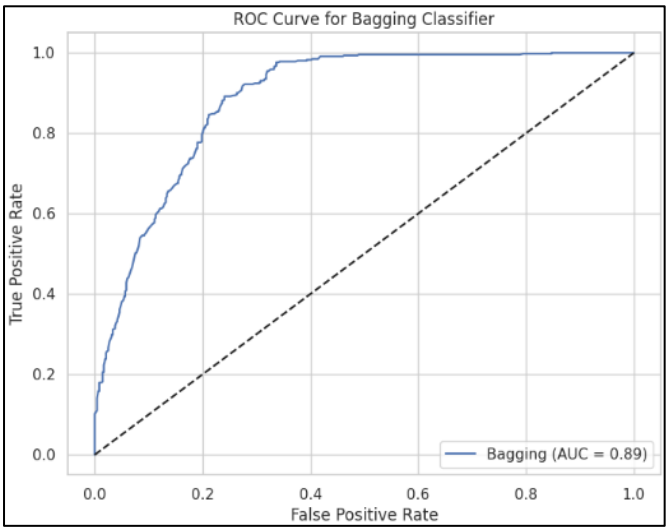


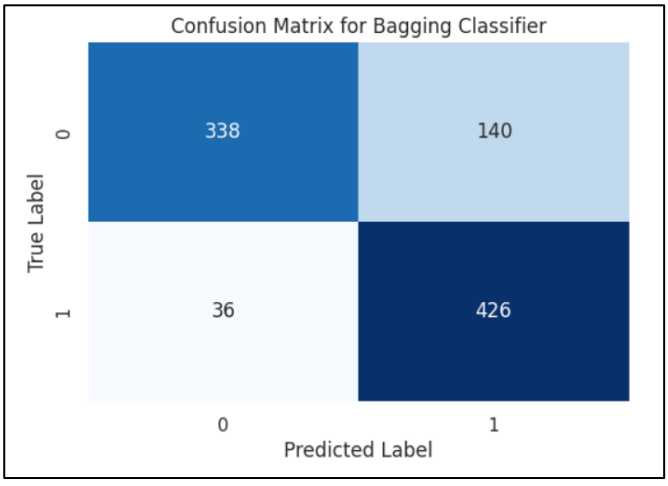
Naïve Bayes:



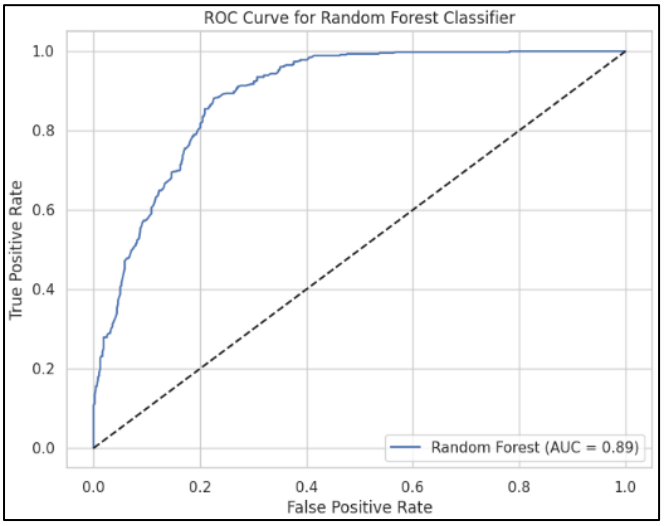


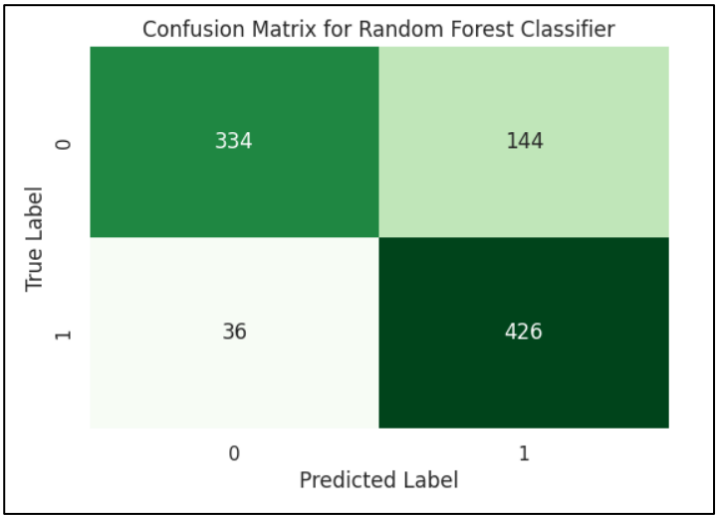
Bagging:

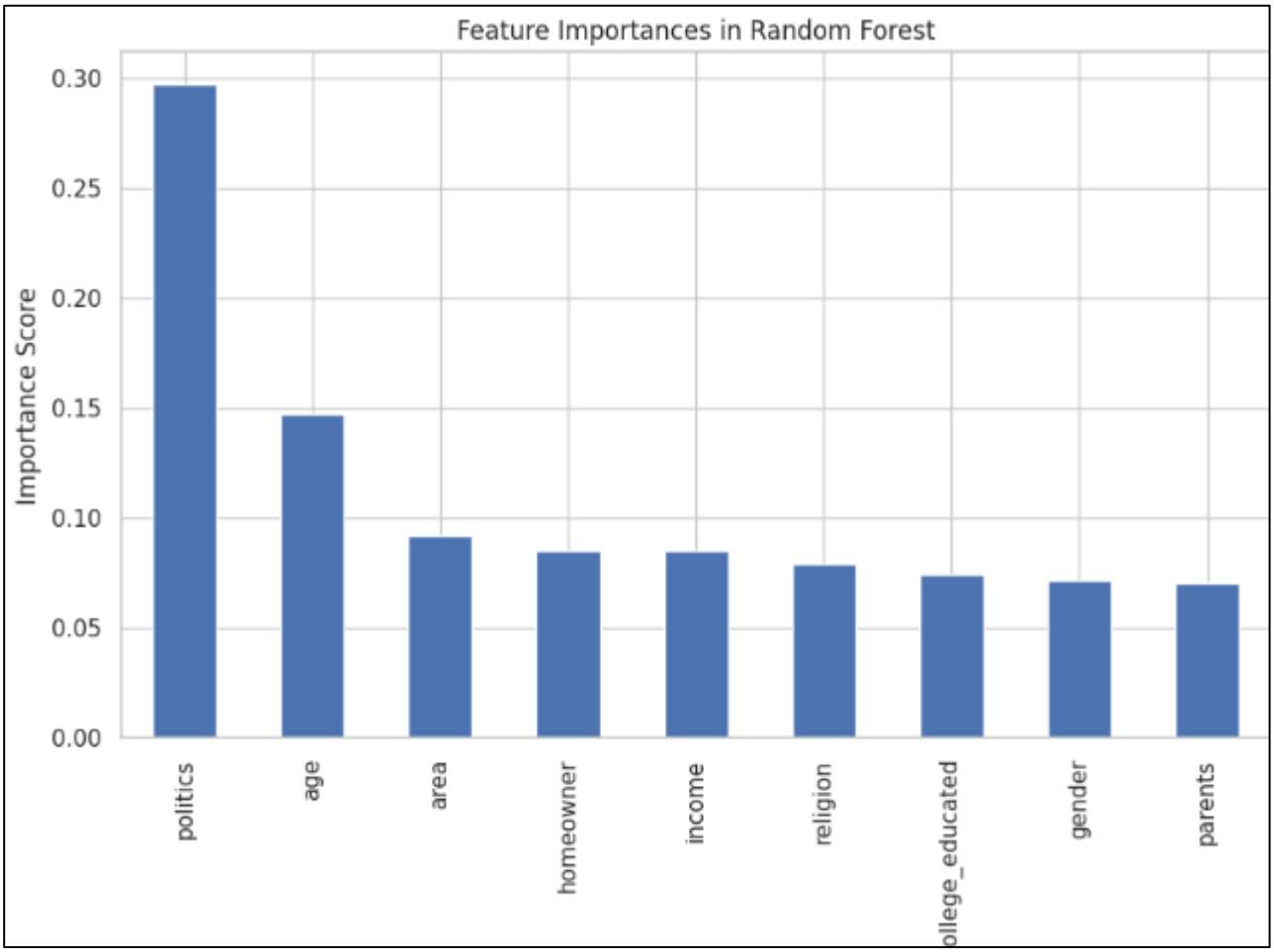




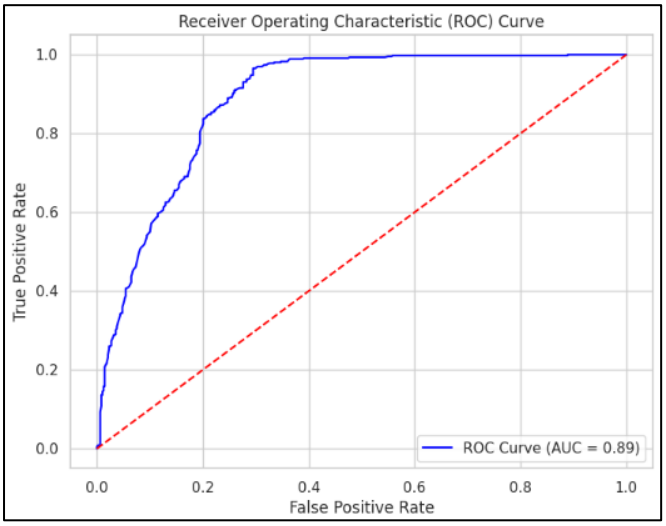
Random Forest:

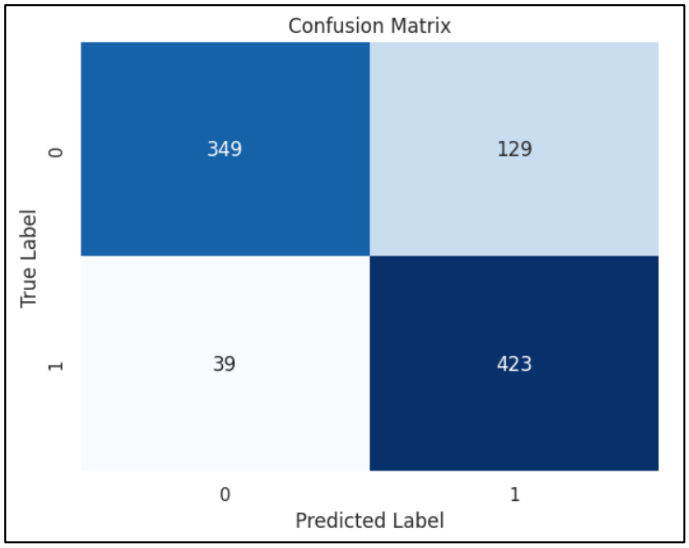




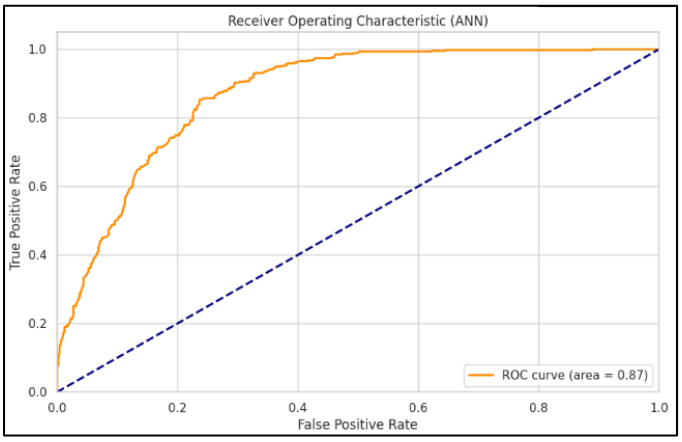


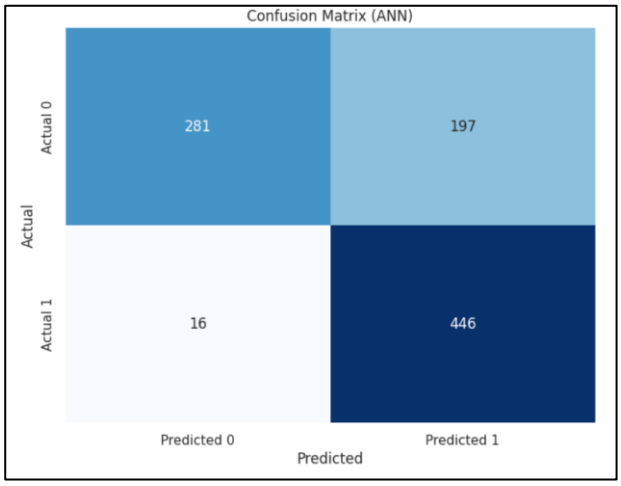
XG Boost:





ANN:





4. Comparison of Performance of Models

Performance Metric	Original Dataset (Data Imbalance)	Balanced Dataset (Post SMOTE)	Balanced Dataset (Post ADASYN)
Decision Tree Classification Model			
ROC	0.55	0.66	0.86
Accuracy	89%	62.23%	80.96%
Error Rate	11%	37.77%	19.04%
Sensitivity / Recall	0%	62.86%	91.71%

Specificity	97%	61.6%	70.46%
Precision	0	59.79%	75.04%
F-Score	0	0.6128	0.8254
Pre-Pruned Decision Tree			
ROC	0.58	0.66	0.86
Accuracy	91.8%	62.23%	80.96%
Error Rate	8.2%	37.77%	19.04%
Sensitivity / Recall	0%	62.86%	91.71%
Specificity	100%	61.6%	70.46%
Precision	-	59.79%	75.04%
F-Score	0	0.6128	0.8254
Post-Pruned Decision Tree			
ROC	0.58	0.66	0.86
Accuracy	91.8%	62.66%	81.17%
Error Rate	8.2%	37.34%	18.83%
Sensitivity / Recall	0%	62.64%	93.07%
Specificity	100%	62.63%	69.62%
Precision	-	60.34%	74.78%
F-Score	0	0.6147	0.8293
Bagging Classifier			
ROC	0.65	0.66	0.89
Accuracy	91.8%	-	81.28%
Error Rate	8.2%	-	18.72%
Sensitivity / Recall	0%	-	92.21%
Specificity	100%	-	70.71%
Precision	-	-	75.26%
F-Score	0	-	0.8288

Random Forest Classifier			
ROC	0.64	0.66	0.89
Accuracy	91.8%	-	80.85%
Error Rate	8.2%	-	19.15%
Sensitivity / Recall	0%	-	92.21%
Specificity	100%	-	69.83%
Precision	-	-	74.74%
F-Score	0	-	0.8256
Naive Bayes Classifier			
ROC	0.64	-	0.65
Accuracy	10.65%	-	60.74%
Error Rate	89.35%	-	39.26%
Sensitivity / Recall	100%	-	65.8%
Specificity	2.79%	-	55.86%
Precision	8.3%	-	59.03%
F-Score	0.153	-	0.6223
Artificial Neural Network			
ROC	0.49	0.66	0.87
Accuracy	92.1%	56.81%	77.34%
Error Rate	7.9%	43.19%	22.66%
Sensitivity / Recall	0%	90.67%	96.54%
Specificity	100%	24.03%	58.79%
Precision	-	53.58%	69.36%
F-Score	0	0.6736	0.8072
XG Boost			
ROC	0.64	0.66	0.89
Accuracy	92.1%	61.81%	82.13%

Error Rate	7.9%	38.19%	17.87%
Sensitivity / Recall	0%	64.72%	91.56%
Specificity	100%	58.91%	73.01%
Precision	-	60.3%	76.63%
F-Score	0	0.6243	0.8343

Inferences from Model Comparison

- XGBoost model has given the best performance when trained with a balanced dataset with regards to all the metrics in consideration
- The performance of all the models is very poor on the original dataset, i.e., the unbalanced dataset. All the models are nearly performing to a random guess as it can be seen through the confusion matrices too.
- The performance of the models has not improved in the case of the dataset developed through the SMOTE algorithm despite establishing a balance in the target class data.
- The performance of the models has been significantly improved through the dataset generated by the ADASYN algorithm
- We can see that the feature importance has changed in predicting the true_conversion.
    - Initially the top 3 features are Income, Religion, and Area and also the score of importance is only around 10%
    - However, in the synthetically developed balanced dataset, the features of importance have changed to Politics, Age, and Area and also the score of importance for politics is around 30%

5. References

- Original Dataset - Ad_Campaign_Dataset
- Dataset after dropping duplicates - Data_No duplicates
- Balanced Dataset (ADASYN) - Balanced Dataset
- Python Notebook - Group 3 ipynb file

THANK YOU