TMall User Conversion Prediction

Team Members: Lingling Deng, Chih-Yun Chang, Yao Chen, Lixing Chen

1. Introduction

Predicting customer conversion is a vital task for e-commerce platforms aiming to maximize sales and improve marketing strategies. This study focuses on data from TMall, a leading e-commerce platform, utilizing K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression algorithms to analyze and predict user conversion rates. By comparing these models, the project aims to identify which algorithm performs better in the TMall context and uncover the most influential features affecting conversion.

The primary objective is to evaluate and compare the predictive performance of KNN, Random Forest, and Logistic Regression in determining whether TMall users will convert. The study also seeks to identify key drivers of conversion, such as campaign spending, user engagement metrics, and website behaviors. The motivation lies in helping TMall optimize its marketing strategies and allocation of resources by leveraging insights from the data.

Challenges

Key challenges include:

- Data Complexity: TMall's dataset includes features like user demographics, campaign characteristics, and website interaction metrics, requiring careful preprocessing and feature selection.
- Imbalanced Data: Conversion events are relatively rare, resulting in an imbalanced dataset that can affect model training.
- Model Trade-offs: Balancing KNN's simplicity, Random Forest's robust performance, and Logistic Regression's interpretability within the context of large-scale e-commerce data.

Report Structure

This report is structured as follows:

1. Introduction: Outlines the importance of predicting customer conversion, key objectives, and challenges in the study.
2. Background: Describes the TMall context, affected stakeholders, and related literature on e-commerce conversion and predictive modeling.
3. Approaches: Details the dataset, feature analysis, and implementation of KNN, Random Forest, and Logistic Regression models.
4. Results and Evaluation: Compares model performance, highlights key predictors, and discusses findings from the analysis.
5. Conclusions: Summarizes key insights, business applications, and future directions for improving conversion predictions.

---

2. Background

TMall Context

TMall, a prominent e-commerce platform in China, serves millions of users daily, offering a wide variety of products. Predicting user conversion is critical for TMall to refine its promotional strategies, allocate marketing budgets effectively, and enhance user experiences. This study leverages TMall's customer data, including campaign characteristics, user demographics, and behavioral metrics, to analyze and predict conversion rates.

Affected Stakeholders

- TMall Marketing Team: Needs insights to design targeted and efficient campaigns.
- TMall Product Teams: Require data-driven strategies to improve customer engagement and satisfaction.
- TMall Executives: Use predictive insights for resource allocation and strategic decision-making.

Literature Review

- E-commerce Conversion: Research on e-commerce platforms like TMall emphasizes the importance of features such as campaign spending, click-through rates, and website interaction in predicting conversions.
- Algorithm Applications: Logistic Regression serves as a baseline model for binary classification, offering interpretability and simplicity. KNN provides an instance-based approach, while Random Forest excels in handling high-dimensional, large-scale data like TMall's.
- Imbalance in Conversion Events: Studies recommend techniques like oversampling or cost-sensitive algorithms to improve model performance on imbalanced datasets common in e-commerce scenarios.

---

3. Approaches

3.1 Data and Data Analysis

Amount of Data

The dataset contains 80,000 rows with 20 features spanning demographics, marketing channels, user engagement metrics, and conversion outcomes. This volume of data is sufficient to explore complex patterns and relationships that cannot be intuited by hand.

Data Analysis and Exploration

The dataset was meticulously analyzed to understand feature distributions, correlations, and their relationship with the target variable, Conversion. Below are key insights:

1. Age Distribution and Conversion Rate (Fig. 1):
    - The customer age ranges from 10 to 70 years, with a relatively uniform distribution.
    - Conversion rates across age groups showed minimal variation, indicating that age does not significantly influence conversion. (Fig. 2)
2. Marketing Channels and Conversion (Fig. 3):

- Five distinct channels (Email, Social Media, PPC, SEO, and Referral) were examined.
- Conversion rates across these channels fluctuated within 2%, suggesting that marketing channels have little impact on conversion.

3. Marketing Type and Conversion (Fig. 4):
   - The marketing types (Awareness, Consideration, Conversion, Retention) directly correlate with the target variable.
   - The Conversion type unsurprisingly shows the highest conversion rate of 93.36%. However, this correlation may not indicate an independent influence on customer behavior but rather reflects the natural progression toward conversion outcomes. Therefore, while useful for understanding overall trends, directly including marketing type as a predictive feature could potentially overstate its impact, as it aligns closely with what is being predicted.
4. Marketing Spend (Fig. 5):
   - Conversion rates showed significant variability with marketing spend, with differences exceeding 10% between spend ranges. (Fig. 6)
   - Higher marketing spending positively correlates with conversion, making it a crucial factor.
5. Click-Through Rate (CTR) (Fig. 7):
   - CTR significantly influenced conversion rates, with fluctuations exceeding 10%. (Fig. 8)
   - Higher CTR values are generally associated with improved conversion rates, emphasizing its importance.
6. Website Visits (Fig. 9):
   - The total number of website visits also impacted conversion rates significantly, with variability exceeding 10%. (Fig. 10)
   - More frequent visits indicate a higher likelihood of conversion, suggesting customer engagement is critical.
7. Time Spent on Site (Fig. 11):
   - The average time per visit showed a noticeable impact on conversion, with variability in conversion rates exceeding 10%. (Fig. 12)
   - Longer time spent on-site correlates with improved conversion rates, highlighting its importance for user engagement.

Correlation Analysis

A heatmap of feature correlations revealed that the following features had the strongest influence on conversion (Fig. 13):

- AdSpend
- ClickThroughRate
- ConversionRate
- WebsiteVisits
- PagesPerVisit
- TimeOnSite
- EmailOpens
- EmailClicks

- PreviousPurchases
- Social Shares
- LoyaltyPoints

These features were prioritized in model development due to their higher predictive power. Features such as AdvertisingPlatform and AdvertisingTool were excluded due to a lack of variability or confidentiality.

3.2 Implementation

1. Objective Functions and Optimization:
    - Target Variable:
      The target variable, Conversion, is binary, where 1 indicates a successful customer conversion, and 0 indicates no conversion. The models are trained to optimize the prediction of this variable, ensuring high accuracy and robust classification.
    - Regularization:
        - Logistic Regression:
          Default L2 regularization is applied to penalize large coefficients, reducing overfitting by simplifying the model.
        - Random Forest:
          Regularization is implicit through ensemble averaging, reducing variance without over-penalizing complexity. Hyperparameters such as the number of trees (n_estimators) and tree depth (max_depth) are tuned to prevent overfitting.
        - KNN:
          No explicit regularization is applied. Instead, the performance is controlled by optimizing the number of neighbors, balancing model complexity and generalization.
    - Class Imbalance:
      The dataset exhibits a class imbalance, with significantly more positive conversions. This is addressed through:
        - Stratified Train-Test Splitting: Ensures proportional representation of both classes in the training and testing datasets.
        - Performance Metrics Beyond Accuracy: Metrics like AUC-ROC, precision, recall, and confusion matrices are prioritized to better evaluate the performance on the minority class.
2. Correctness:

    Each model was tuned and evaluated to ensure it appropriately models the conversion prediction task without overfitting or underfitting.

    1. K-Nearest Neighbors (KNN):
        - Hyperparameter Tuning: The number of neighbors (K) was tested across a range of values (1–20) to identify the optimal value. The best performance was observed at K=16 (Fig. 14)
        - Performance:
            - Test Accuracy: 87.68%.

        - AUC: 0.58, indicating poor discrimination between classes.
    - Observations:
        - Struggles with high-dimensional data and imbalanced classes, as seen in the low AUC score.
        - It underfits the minority class due to its reliance on Euclidean distance.
2. Logistic Regression:
    - Optimization:
        - L2 regularization mitigates overfitting, ensuring a balance between bias and variance.
        - A maximum iteration cap (max_iter=1000) ensures convergence for large datasets.
    - Performance:
        - Test Accuracy: 87.62%.
        - AUC: 0.74, reflecting moderate discriminatory power.
    - Observations:
        - Provides interpretability and robustness.
        - Consistently underpredicts the minority class, favoring majority class predictions.
3. Random Forest:
    - Optimization:
        - Hyperparameters such as the number of trees (n_estimators=100), maximum depth (max_depth), and minimum samples per split (min_samples_split) were tuned.
        - Out-of-bag (OOB) evaluation was used to validate performance and detect overfitting.
    - Performance:
        - Test Accuracy: 89.56%.
        - AUC: 0.81, the best among tested models.
    - Observations:
        - Training accuracy of 100% indicates slight overfitting.
        - Feature importance analysis provided actionable insights, with key features like AdSpend, ClickThroughRate, and PagesPerVisit driving predictions.

---

4 Results and Evaluation

1. Evaluation Metrics:
    - Accuracy: While useful, accuracy is insufficient in class-imbalanced datasets.
    - AUC-ROC: Measures the ability of the model to distinguish between classes across thresholds.
    - Confusion Matrix: Highlights the distribution of predictions across actual classes, particularly false positives and negatives.

2. Model Performance:
    - KNN:
        - Test Accuracy: 87.68%.
        - Confusion Matrix indicates the model performs poorly on the minority class. (Fig. 15)
        - AUC (0.58) suggests weak discrimination capability. (Fig. 16)
    - Logistic Regression:
        - Test Accuracy: 87.62%.
        - Confusion Matrix shows consistent underprediction of the negative class. (Fig. 17)
        - AUC (0.74) indicates moderate discrimination. (Fig. 18)
    - Random Forest:
        - Test Accuracy: 89.56%.
        - Confusion Matrix shows significant improvement in identifying the minority class. (Fig. 19)
        - AUC (0.81) suggests the best performance among the models tested. (Fig. 20)
3. Model Insights:
    - Feature importance analysis (Random Forest) revealed that variables like AdSpend, ClickThroughRate, and PreviousPurchases are critical predictors of conversion. (Fig. 21)
    - Stability of Random Forest was assessed using bootstrapped ROC curves, confirming consistent performance with narrow confidence intervals. (Fig. 22)
4. Confidence in Outcomes:

    The Random Forest model demonstrates the best trade-off between performance and generalizability. Among the tested models, Random Forest outperformed others with a test accuracy of 89.56% and high reliability, making it the most suitable choice for predicting user conversions on TMall. Logistic Regression offered interpretability, while KNN struggled with the dataset complexity.

---

5 Conclusions

Key Learnings:

1. Impactful Features: The study identified critical factors influencing customer conversion, such as AdSpend, TimeOnSite, PagesPerVisit, and ClickThroughRate. These insights emphasize the importance of user engagement and effective marketing strategies.
2. Model Performance: Among the tested models, Random Forest outperformed others with a test accuracy of 89.56% and high reliability, making it the most suitable choice for predicting user conversions on TMall.

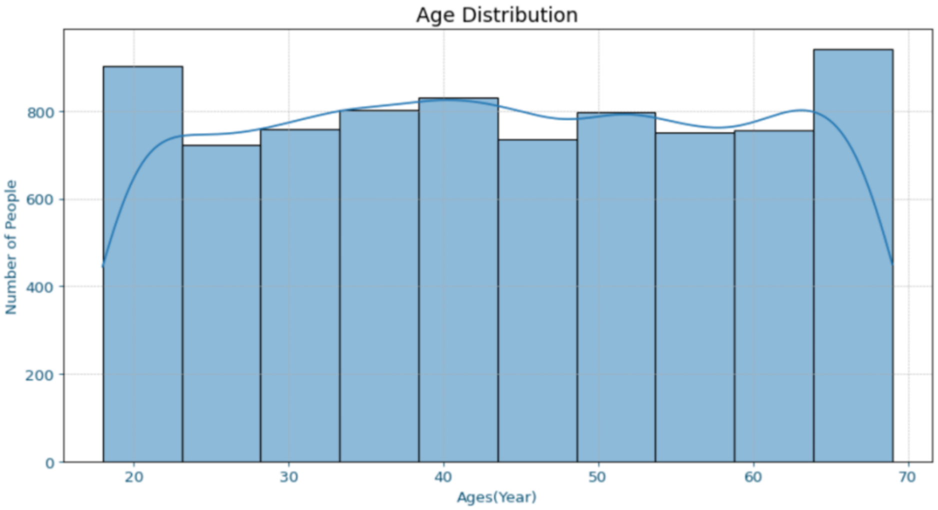
3. Business Applications: The findings can directly support TMall’s marketing optimization by prioritizing high-impact features and tailoring campaigns to improve user engagement and conversion rates.

Future Directions:
Enhancements like real-time prediction systems, advanced algorithms (e.g., XGBoost), and richer feature sets (e.g., customer sentiment) could further refine the model and its applications for TMall’s strategic growth.
This project underscores the value of data-driven decision-making in e-commerce, providing actionable insights to optimize marketing strategies and improve conversion outcomes.

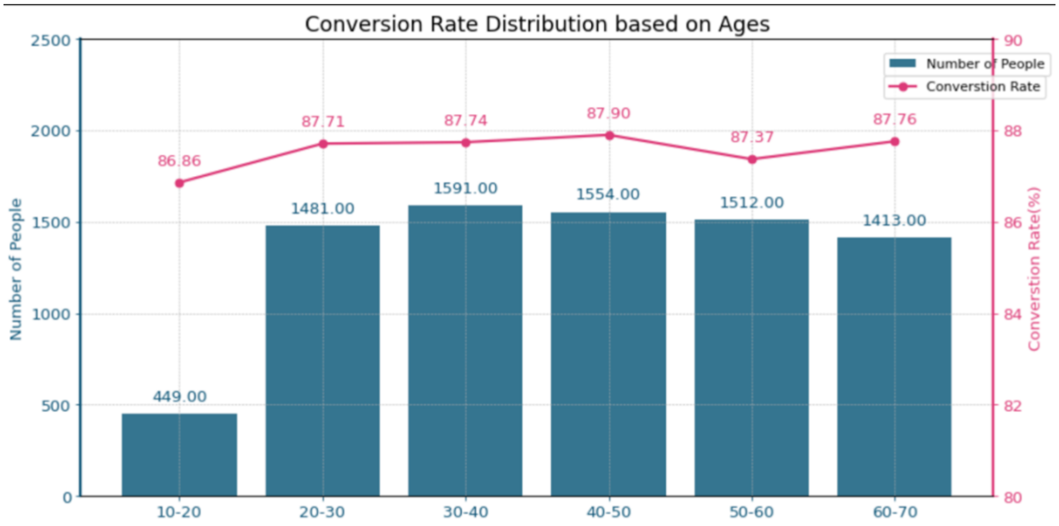
---

6 Figures

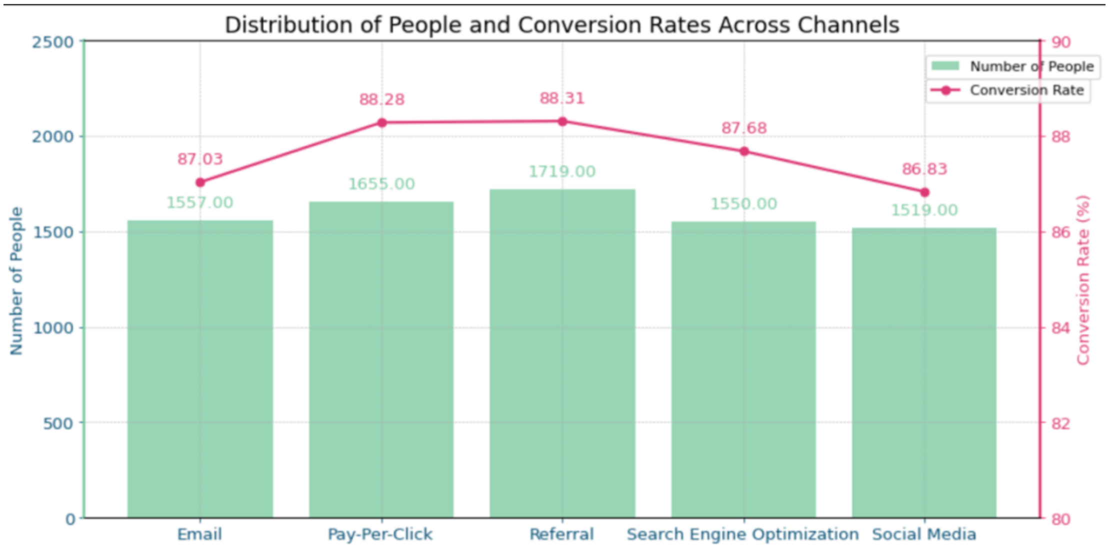
- **Fig. 1: Age Distribution**



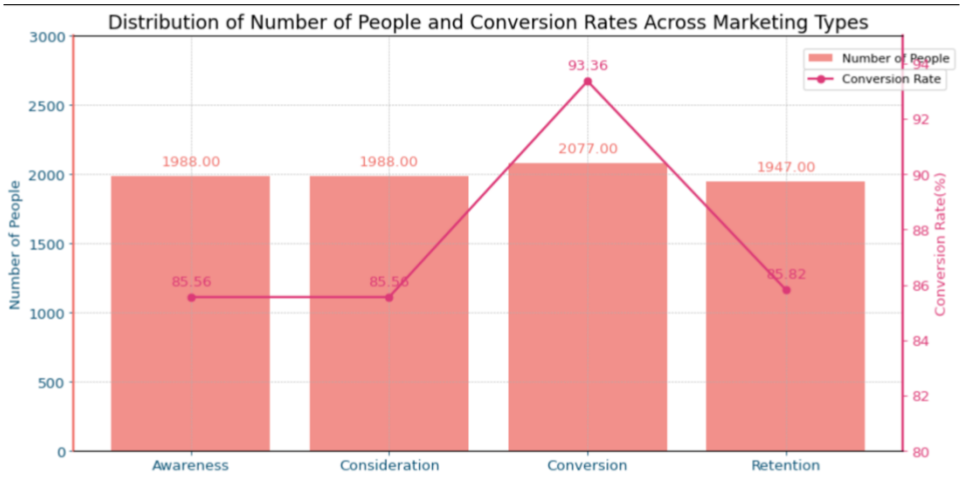
- **Fig. 2: Conversion Rate Distribution based on Ages**



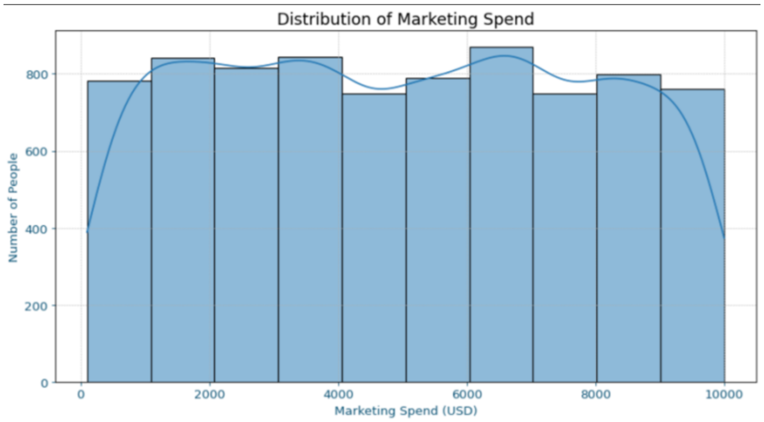
- **Fig. 3: Distribution of People and Conversion Rates Across Channels**



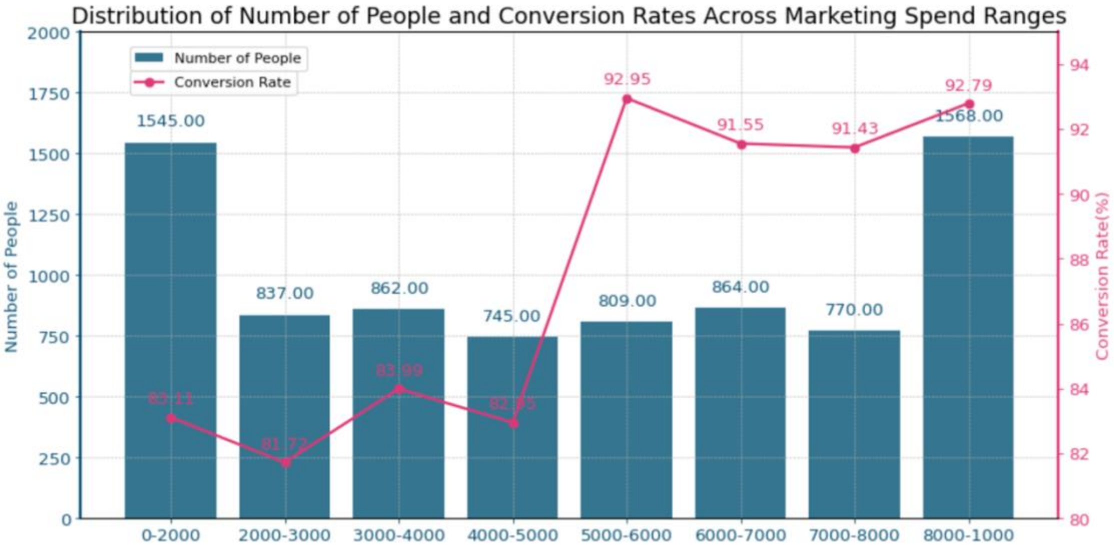
- **Fig. 4: Distribution of Number of People and Conversion Rates Across Marketing Types**



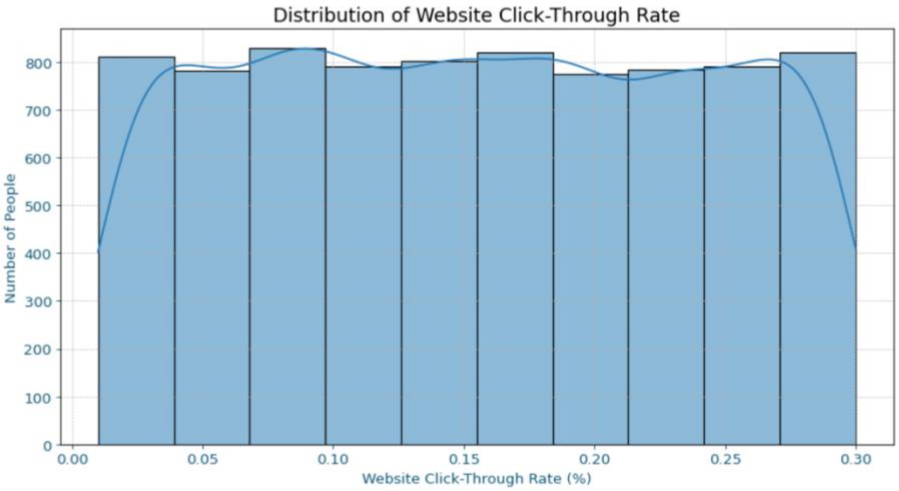
- **Fig. 5: Distribution of Marketing Spend**



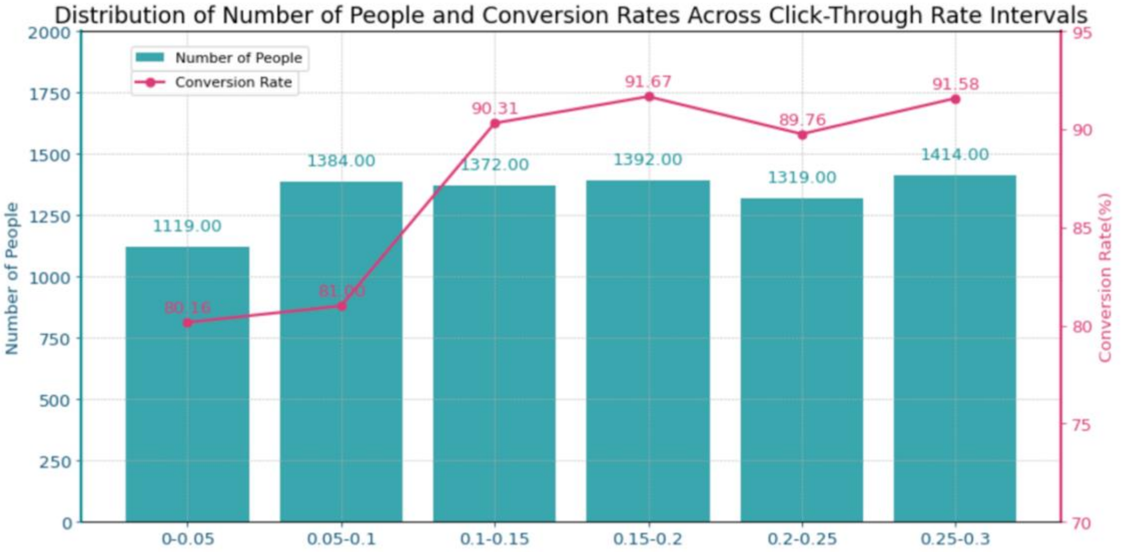
- **Fig. 6: Distribution of Number of People and Conversion Rates Across Marketing Spend Ranges**



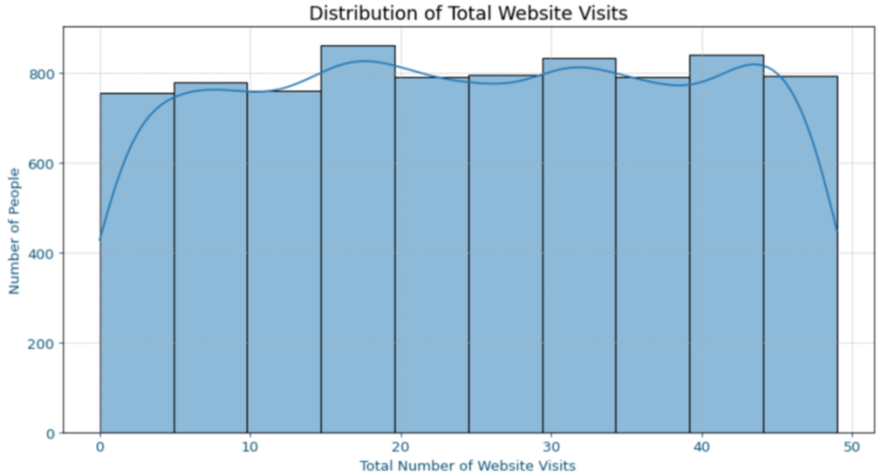
- **Fig. 7: Distribution of Website Click-Through Rate**



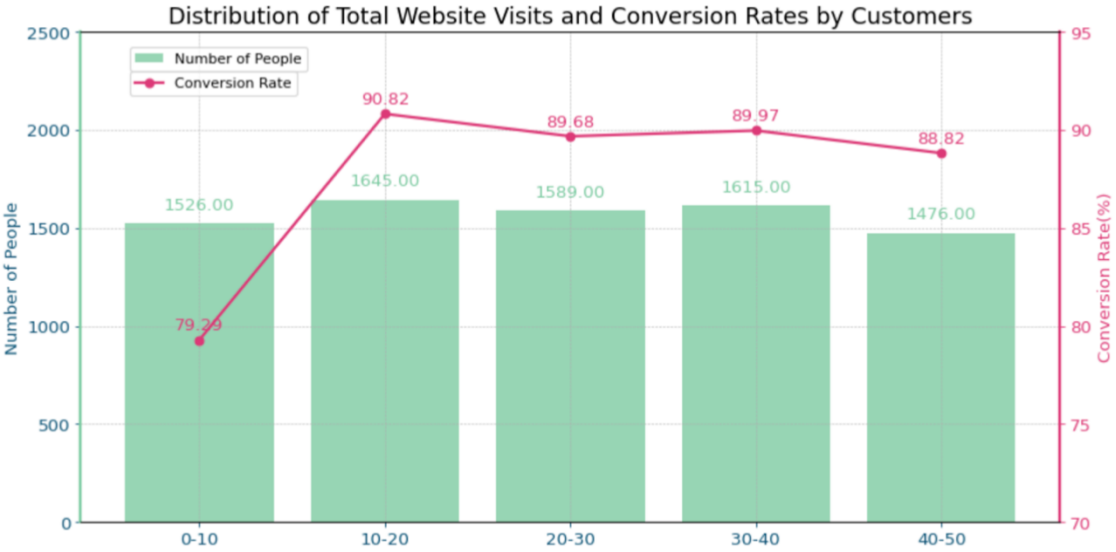
- **Fig. 8: Distribution of Number of People and Conversion Rates Across Click-Through Rate Intervals**



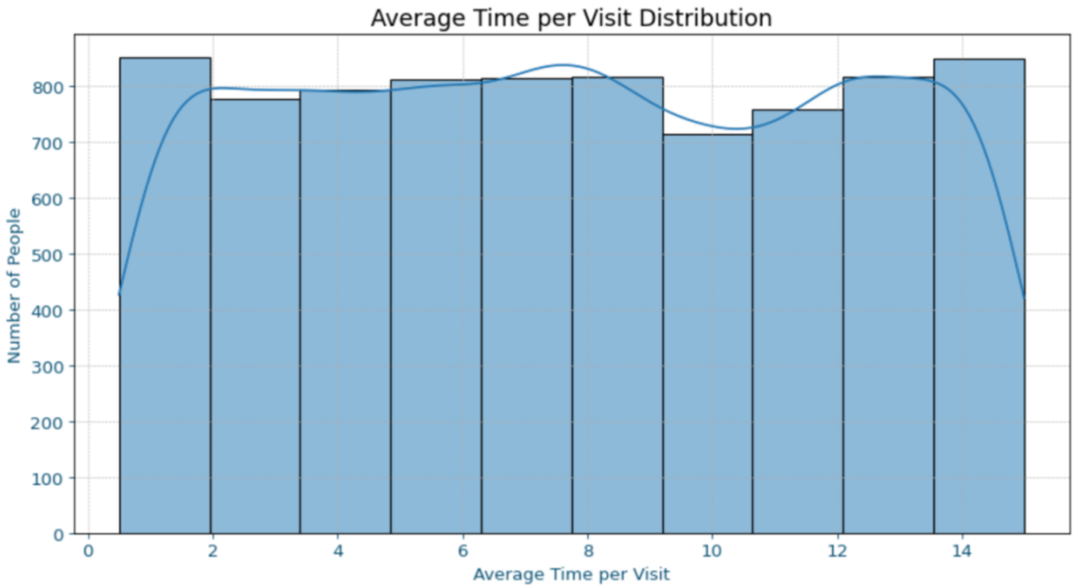
- **Fig. 9: Distribution of Total Website Visits**



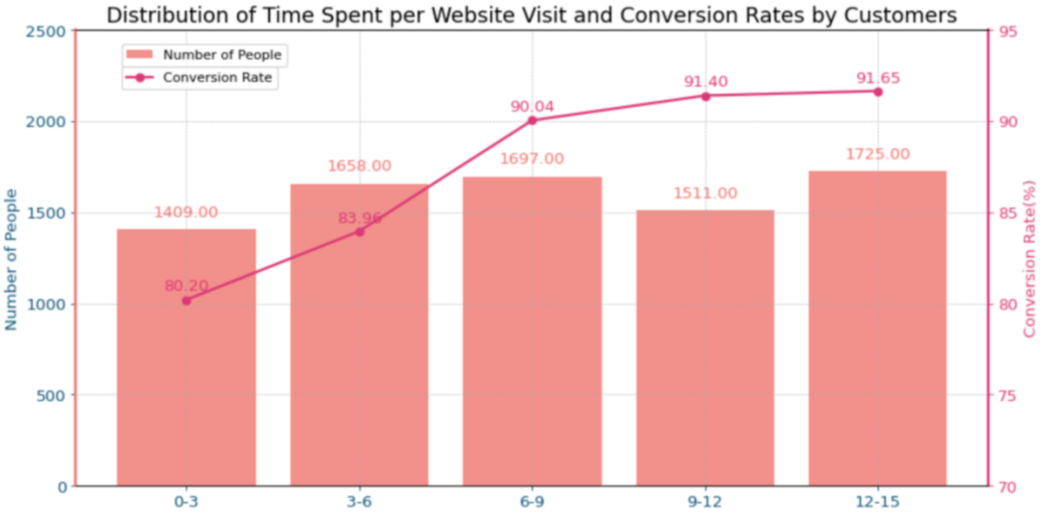
- **Fig. 10: Distribution of Total Website Visits and Conversion Rates by Customers**



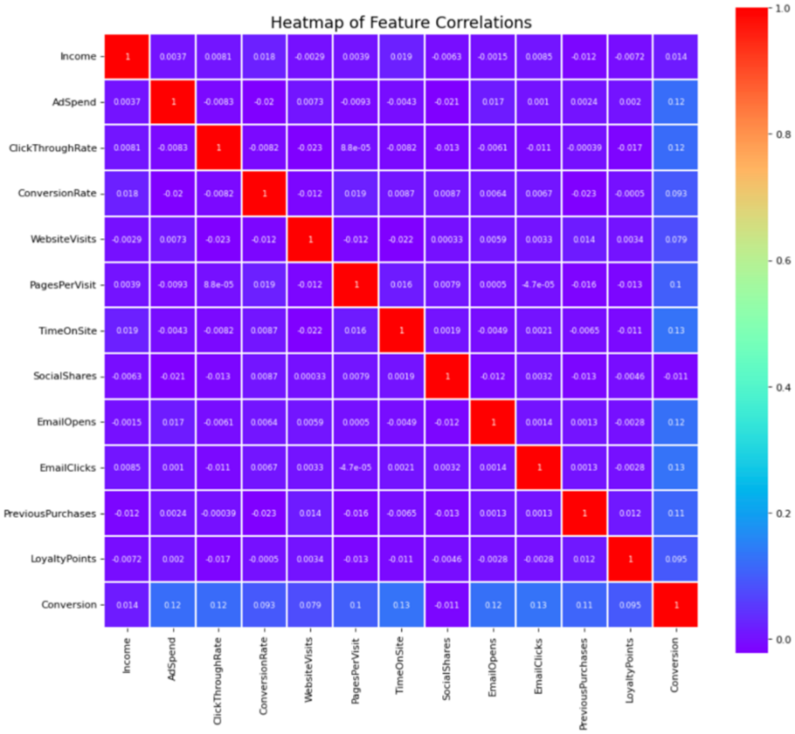
- **Fig. 11: Average Time per Visit Distribution**



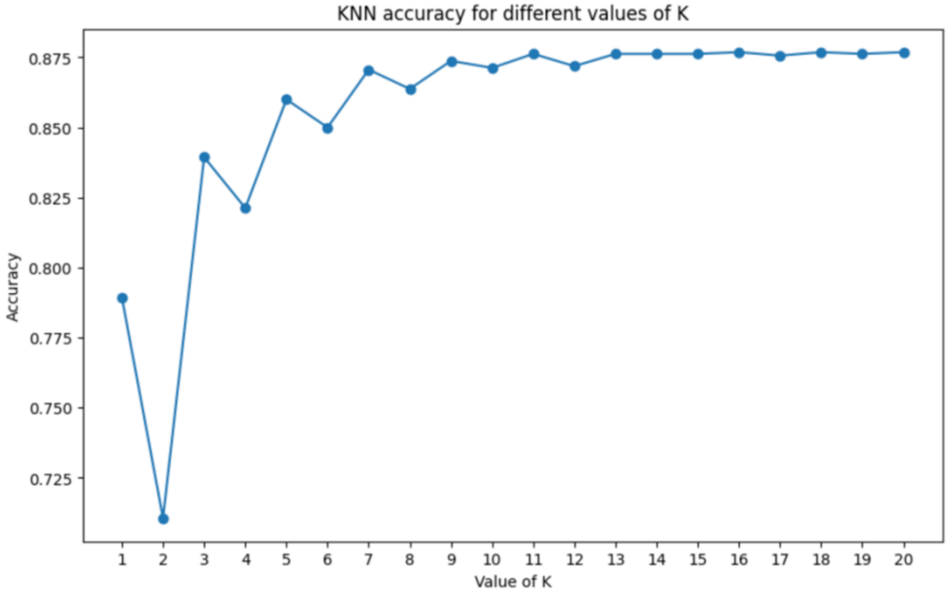
- **Fig. 12: Distribution of Time Spent per Website Visit and Conversion Rates by Customers**



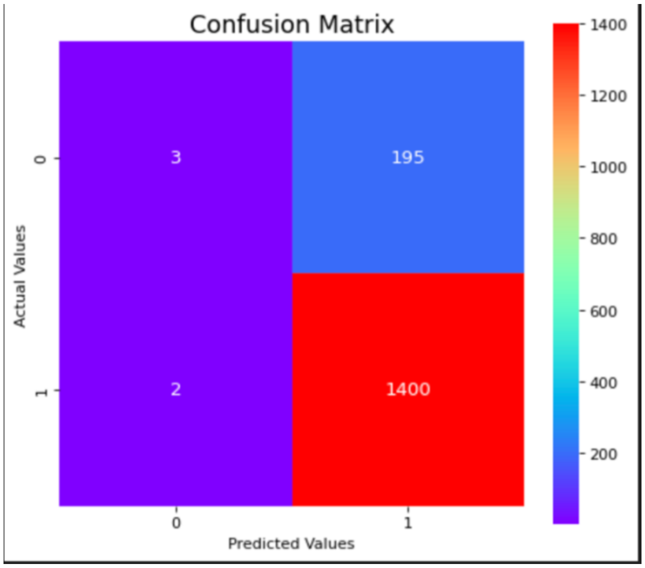
- **Fig. 13: Heat Map Feature Correlation**



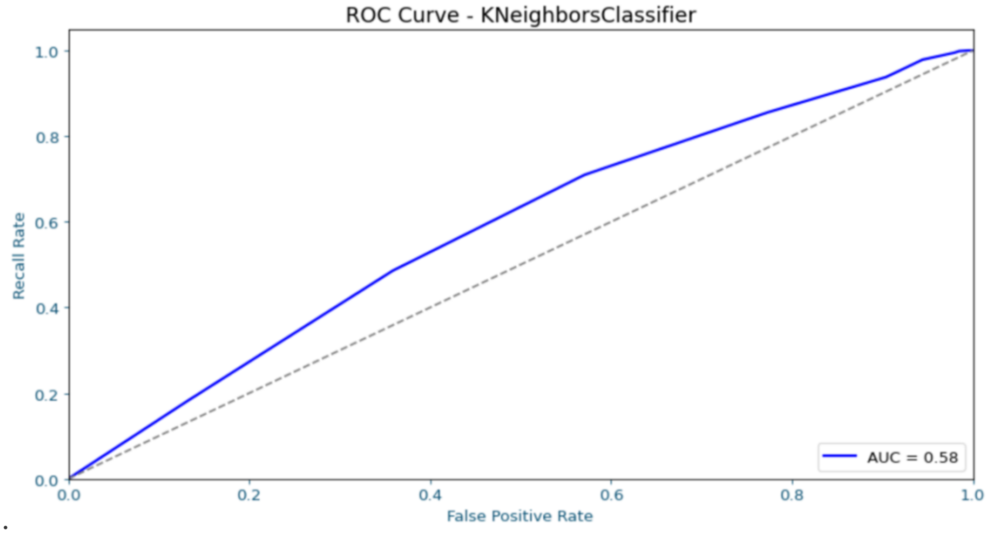
- **Fig. 14: KNN Accuracy for Different Values of K**



- **Fig. 15: Confusion Matrix for KNN**

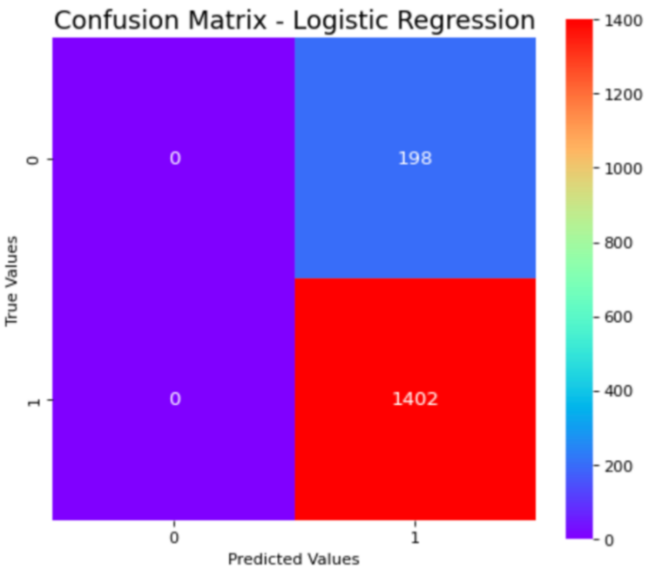


- **Fig. 16: ROC Curve for KNN**

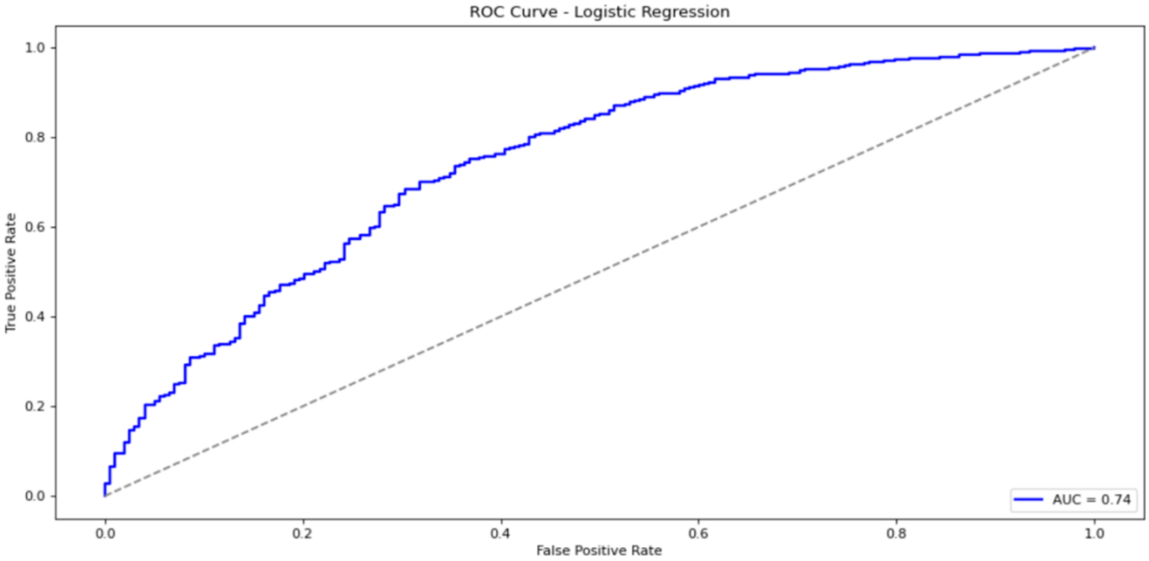


.

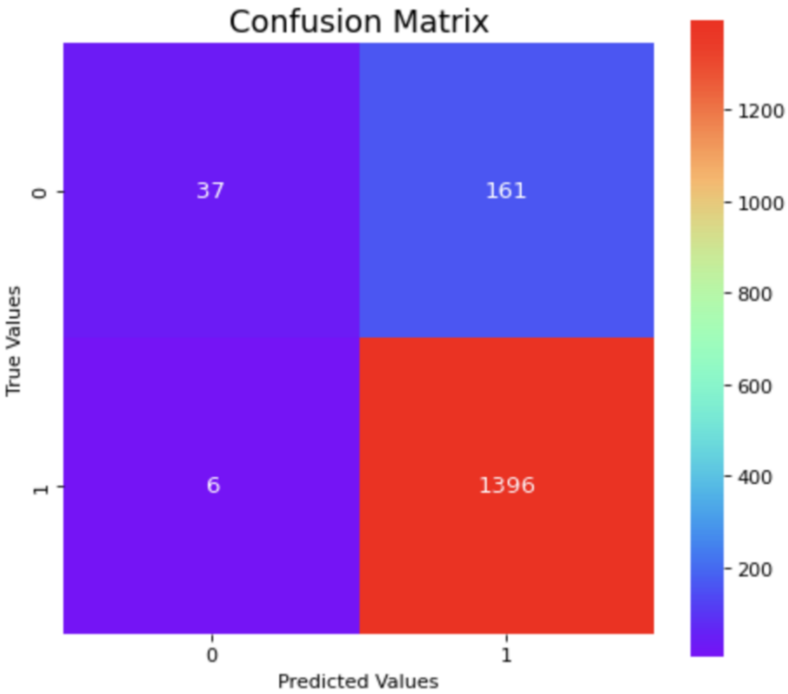
- **Fig. 17: Confusion Matrix - Logistic Regression**



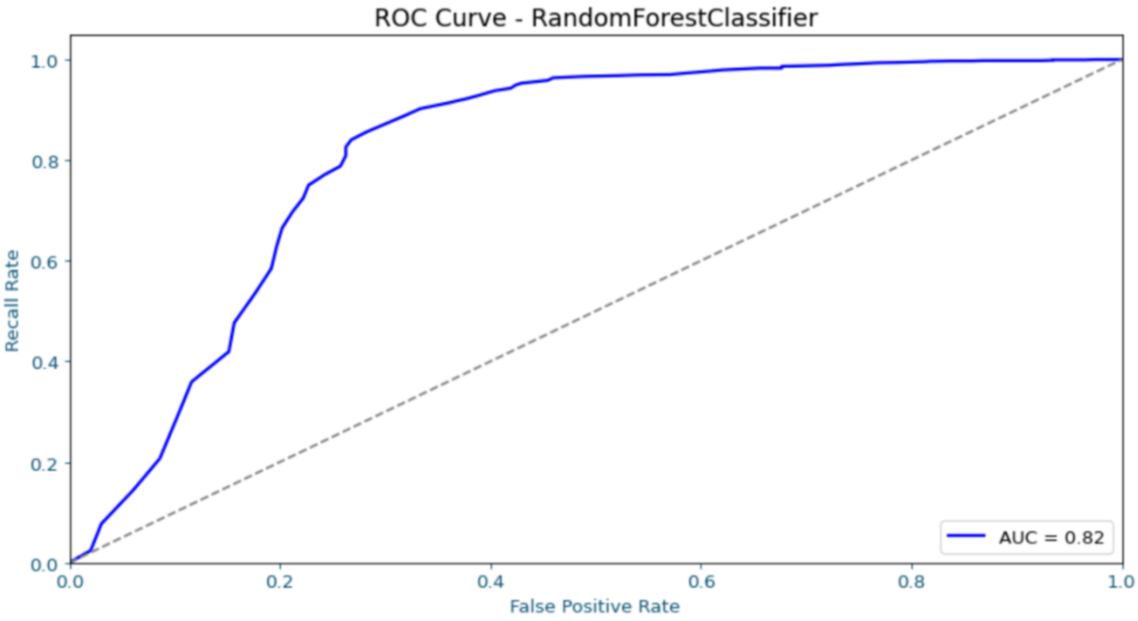
- **Fig. 18: ROC Curve - Logistic Regression**



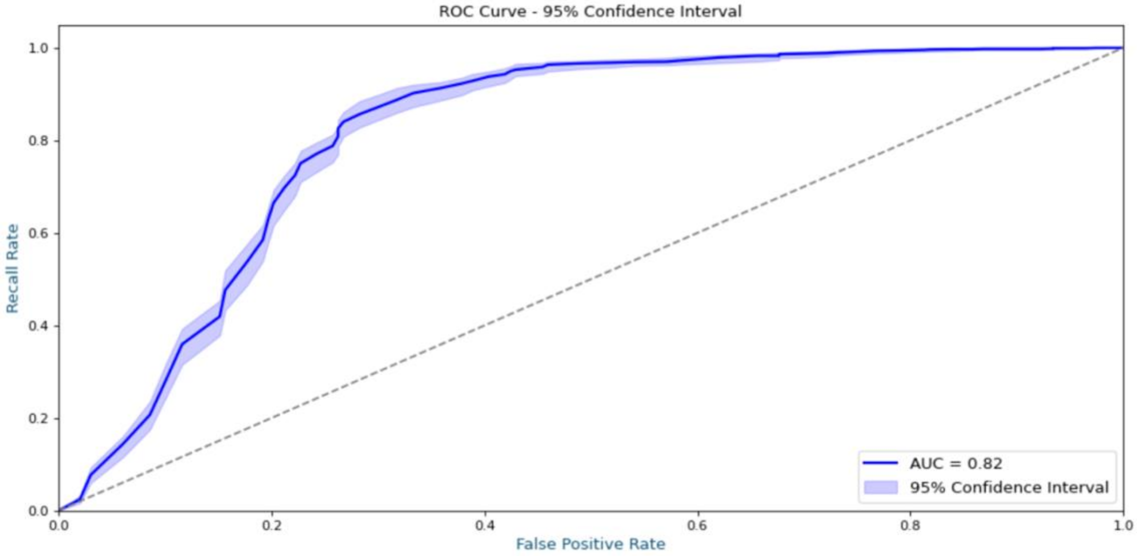
- **Fig. 19: Confusion Matrix - Random Forest**



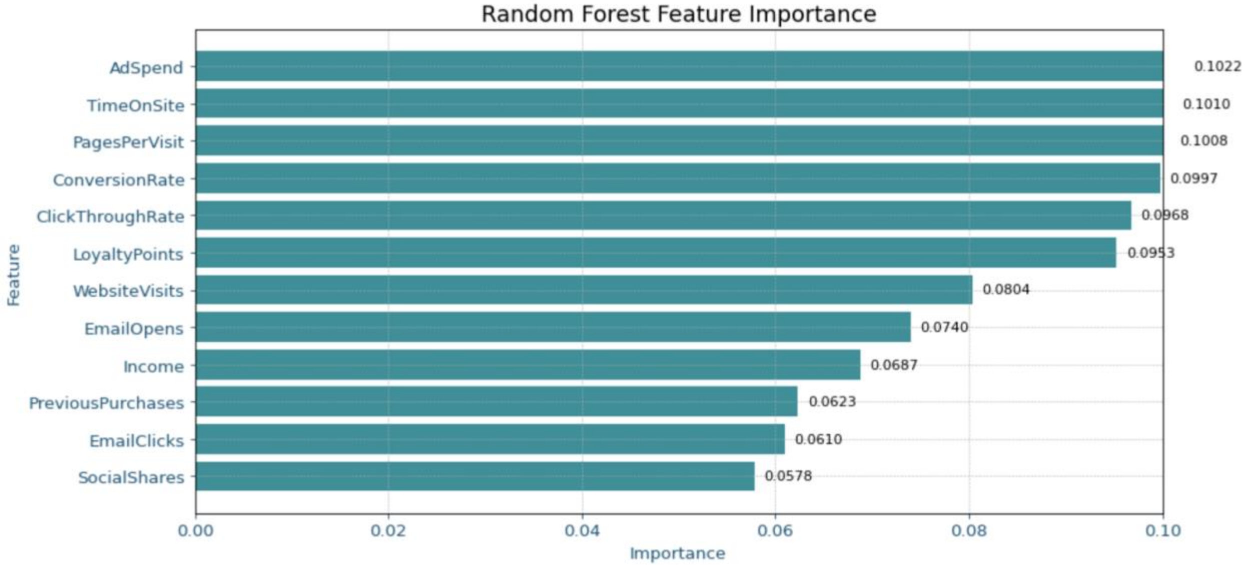
- **Fig. 20: ROC Curve - Random Forest**



- **Fig. 21: ROC Curve - Random Forest with 95% Confidence Interval**



- **Fig. 22: Random Forest Feature Importance**


Random Forest Feature Importance
Feature
AdSpend
TimeOnSite
PagesPerVisit
ConversionRate
ClickThroughRate
LoyaltyPoints
WebsiteVisits
EmailOpens
Income
PreviousPurchases
EmailClicks
SocialShares
Importance