Bayesian Analysis iFood CRM Data Analyst Case

Authors

Chetan Reddy Valluru, Chittha Bhuvan Reddy Diddekunta, Nancy Soni

Statistical Modeling & Computing - 16:954:567:01 , Rutgers University - NB

ABSTRACT

This report presents an analysis of customer data and the development of predictive models aimed at optimizing iFood's CRM campaigns. The objective was to understand customer behavior, segment customers based on their characteristics, and build predictive models to maximize the profitability of future marketing campaigns. The model utilizes logistic regression and incorporates socio-demographic data, past purchase behavior, and interaction metrics to make predictions. Through comprehensive analysis and evaluation, insights into model performance and predictive features are provided.

INTRODUCTION

In today's competitive market landscape, effective marketing campaigns are essential for businesses to engage customers and drive revenue growth. Predictive modeling techniques offer valuable insights into customer behavior and enable targeted marketing strategies. This report focuses on the development and evaluation of predictive models for optimizing iFood's CRM campaigns. By leveraging logistic regression and incorporating socio-demographic data, past purchase behavior, and interaction metrics, insights into model performance and predictive features are provided to enhance campaign effectiveness and drive business growth.

The objectives of this analysis are threefold:

1. Explore the dataset to gain a comprehensive understanding of customer characteristics and behaviour.
2. Propose a segmentation strategy based on customer behaviours to enable targeted marketing efforts.
3. Develop predictive models to forecast customer response to future marketing campaigns and maximize profitability.

Through the application of advanced analytics techniques, including exploratory data analysis, customer segmentation, and predictive modelling, this report aims to provide actionable recommendations to the marketing team at iFood, ultimately driving more effective and efficient marketing campaigns.

DATASET

The dataset provided contains socio-demographic and firmographic features of 2,240 customers who were contacted for a pilot CRM campaign. The dataset includes information such as age, income level, previous purchase history, and engagement with previous marketing initiatives. Additionally, a flag indicates whether each customer responded to the campaign by purchasing the promoted product.

Key features of the dataset include:

- Socio-demographic information: Age, income level, geographic location.
- Firmographic features: Previous purchase history, engagement with previous marketing campaigns.
- Response indicator: Flag indicating whether the customer responded to the campaign.

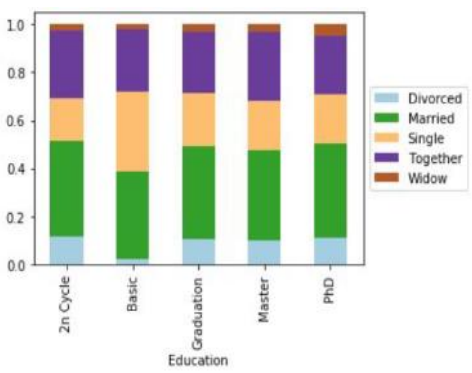
Feature	Description
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if customer accepted the offer in the last campaign, 0 otherwise
Complain	1 if customer complained in last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on gold products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's website
NumWebVisitsMonth	number of visits to company's website in the last month
Recency	number of days since the last purchase

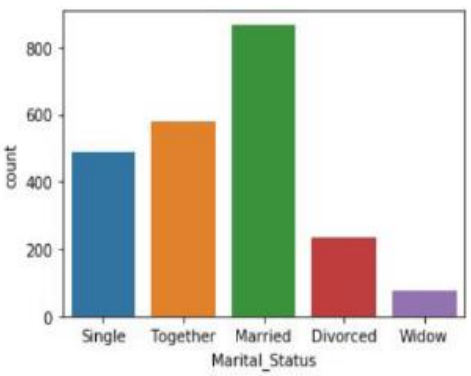
DATA EXPLORATION

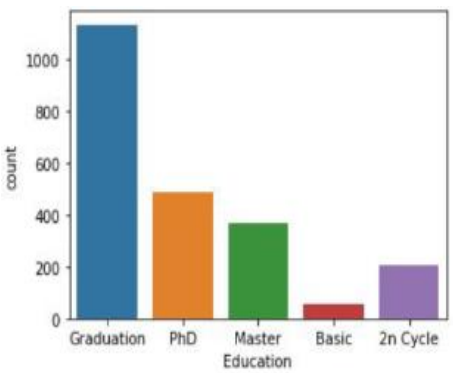
We delve into the dataset to gain insights into the socio-demographic and firmographic characteristics of iFood's customers. The analysis includes graphical visualizations and correlation analysis to better understand customer profiles and behaviors.

Customer Profile Analysis

Below are some graphical visualizations of the customers distributions. The majority of customers exhibit a high level of education and are married. However, there is no significant correlation observed between education and marital status.



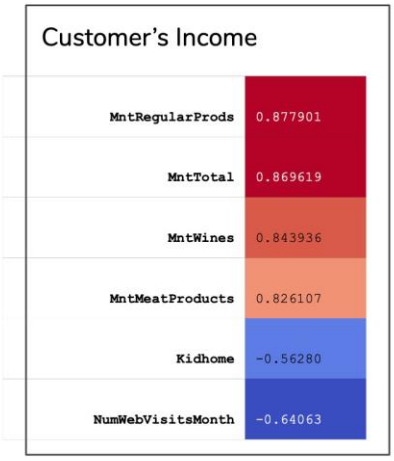




- A clear trend is observed where total spending increases proportionally with customers' incomes.
- This suggests that higher-income customers tend to spend more on iFood's offerings.

Correlation analysis

1. **Income as a Proxy for Various Features:**
   - Income serves as a proxy for several other features, including the amount spent.
   - Positive correlations are found between income and expenditure on meat and wine, while a negative correlation exists with visits to the company's website.
2. **Relationship between Wine Expenditure and Income:**
   - Expenditure on wine is not only associated with high income but also correlates with spending on meat and the purchasing channels (catalog, stores, or website).
3. **Effect of Number of Kids on Income and Spending:**
   - The number of kids in a household shows a negative correlation with income and total spending.
   - Specifically, households with more children tend to spend less and are less likely to purchase wine.
4. **Income and Acceptance of Campaigns:**
   - Higher income levels are positively associated with accepting marketing campaigns, indicating a potential target segment for future campaigns.

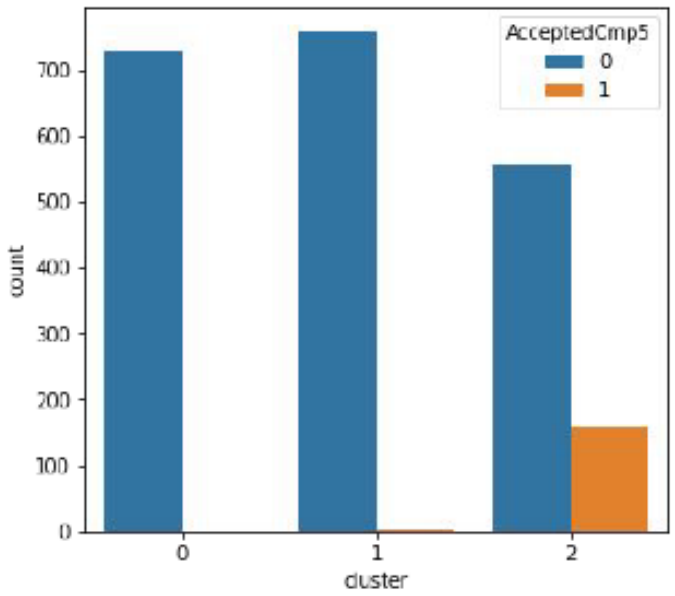


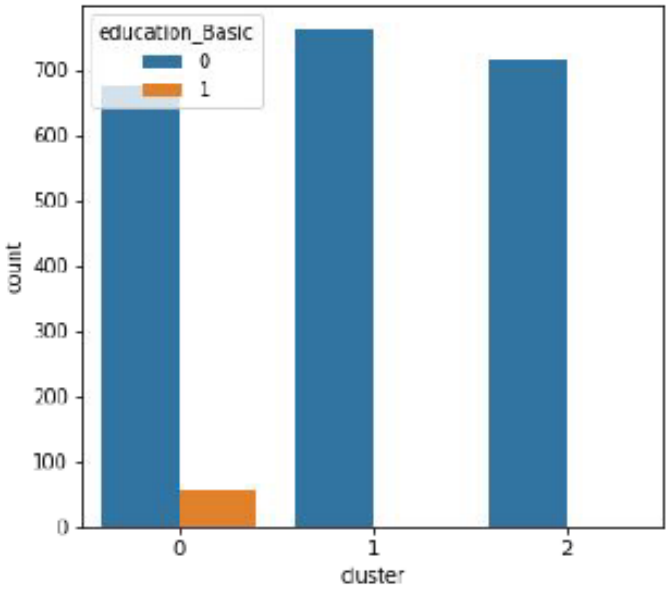
Customer Segmentation

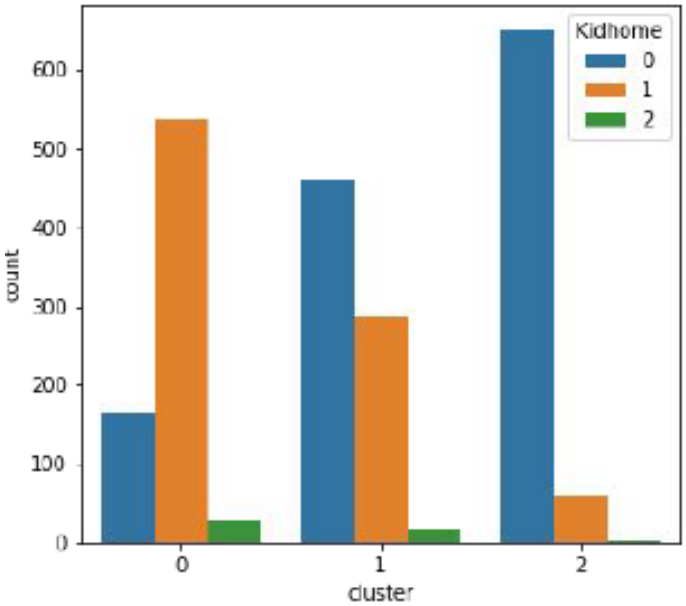
Based on income levels, customers were segmented into three categories: low, average, and high income. The analysis revealed distinct characteristics within each segment:

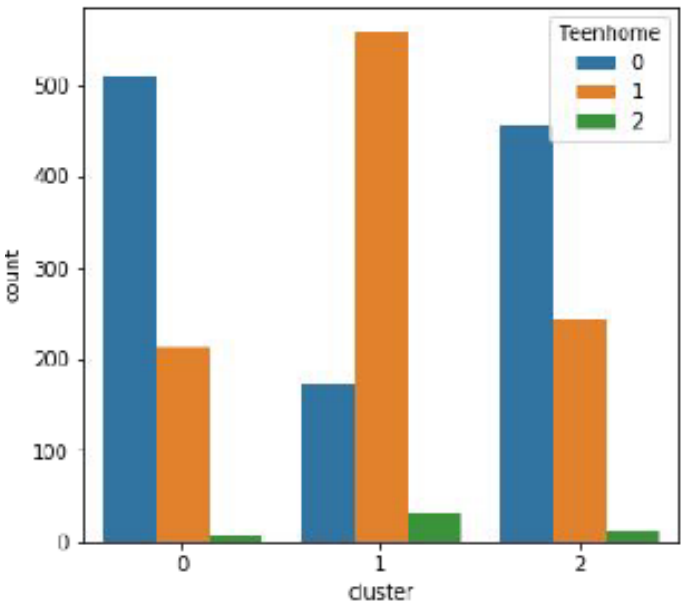
- Low-Income Segment: Predominantly consists of individuals with one child and no teenagers.
- Average-Income Segment: Dominated by customers without children but with one teenager in their household.
- High-Income Segment: Mainly comprises individuals without children or teenagers.

These findings provide valuable insights for targeted marketing strategies tailored to each income segment, aiming to maximize campaign effectiveness and customer engagement.









MODEL DEVELOPMENT

In this study, a logistic regression model was developed to forecast customer responses to marketing campaigns, leveraging features such as age, income, past purchase history, and interaction metrics. The methodology began with the preprocessing of training and validation datasets, where missing income values were imputed using the median to maintain data integrity. Features were carefully selected based on their potential influence on customer behavior, and data was scaled using Standard Scaler to ensure uniform contribution to the model's predictions. The logistic regression model was then trained using the scaled training data.

Model performance was evaluated using the validation dataset with multiple metrics to ensure comprehensive assessment. Key evaluation metrics included the accuracy score, confusion matrix, classification report, and ROC and Precision-Recall curves. These metrics provided insights into the model's ability to predict accurately and understand the trade-offs between various aspects of its performance. The results from these evaluations highlighted the model's effectiveness in identifying customer segments and optimizing future CRM campaigns, illustrating its practical implications for targeted marketing strategies.

RESULTS

The logistic regression model achieved a commendable accuracy score of **0.9048473967684022** in predicting customer response to the marketing campaign. The confusion matrix revealed **495 True Negatives, 14 False Positives, 39 False Negatives, 9 True Positives** indicating strong performance in correctly classifying both positive and negative responses. The classification report provided further insights, with precision, recall, and F1-score metrics indicating [insert insights from classification report analysis.

Visualizations such as the ROC curve and Precision-Recall curve illustrated the model's performance trade-offs. The ROC curve demonstrated With an AUC of **0.89** with the area under the curve (AUC) indicating the overall discriminatory power of the model. Similarly, the Precision-Recall curve highlighted AP of **0.41**, providing insights into the precision-recall trade-off and optimal threshold selection.

Overall, the results suggest that the logistic regression model effectively predicts customer response to the marketing campaign. The model's accuracy and performance metrics provide valuable insights into customer behavior and enable targeted marketing strategies to enhance campaign effectiveness.

EXPERIMENTATION WITH BAYESIAN ANALYSIS

Bayesian analysis incorporates prior knowledge through the use of prior distributions and Bayesian inference. In the context of predicting customer responses, Bayesian logistic regression can be particularly useful

1. Setting Priors: Based on historical data from past marketing campaigns, especially those that failed, priors can be set for the logistic regression coefficients. These priors reflect the beliefs about the effects of various features before observing the current campaign data.

2. Bayesian Model Fitting: Unlike traditional logistic regression, Bayesian logistic regression computes the posterior distribution of the coefficients given the data, rather than finding point estimates. This involves updating the priors in the light of new data, typically using computational techniques such as Markov Chain Monte Carlo (MCMC) methods.

3. Posterior Analysis and Prediction: The posterior distributions of the model parameters are used to make probabilistic predictions about whether a customer will respond to a new campaign. This approach not only provides predictions but also quantifies uncertainty in these predictions, which can be crucial for making informed decisions under uncertainty.

4. Comparison with Non-Bayesian Models: The performance of the Bayesian model can be compared with the non-Bayesian logistic regression to assess the value of incorporating prior knowledge into the model.

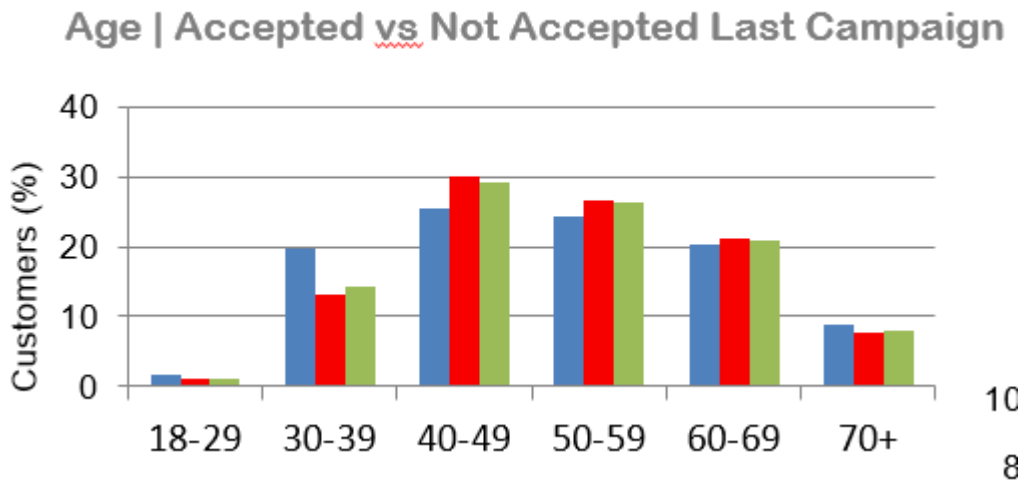
This detailed approach in Bayesian analysis allows for a deeper understanding of model behavior and offers a robust framework for predicting customer responses by effectively integrating historical campaign data.

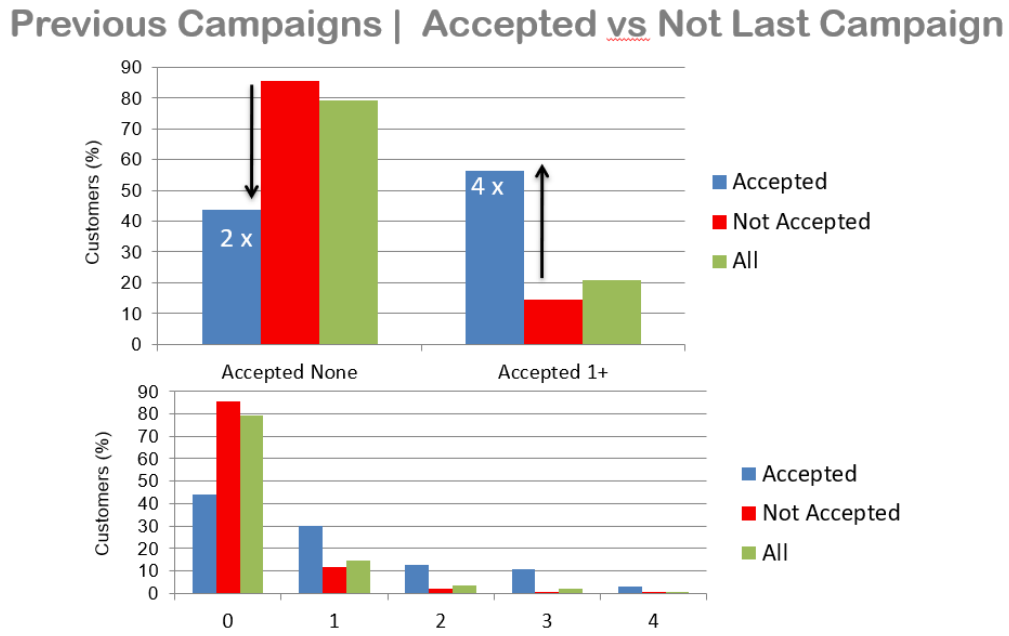
LITERATURE CITED

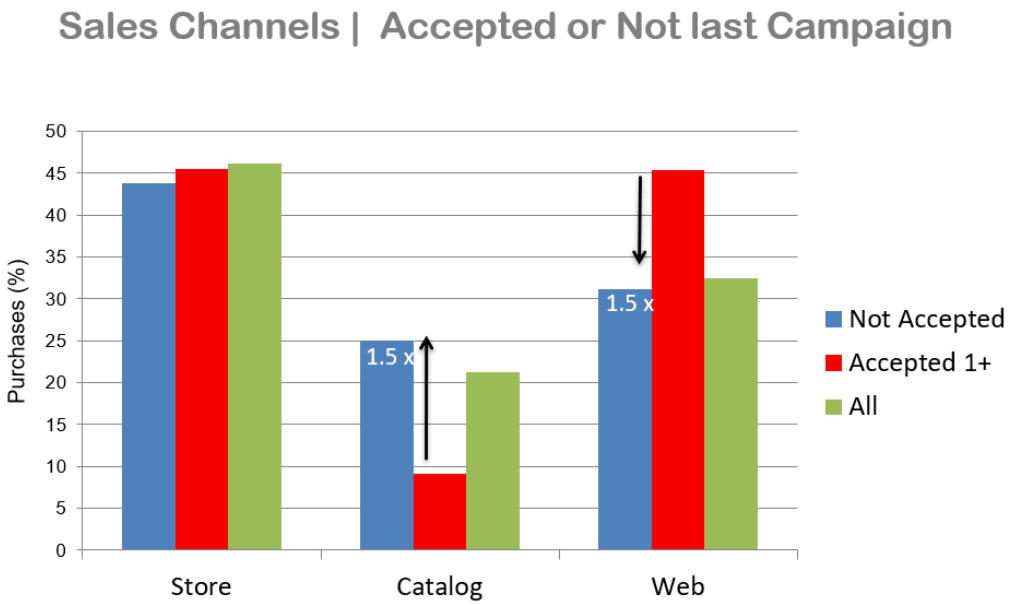
1. https://medium.com/accelerated-analyst/how-to-use-logistic-regression-for-practical-business-problems-e4582826a802 - Practical Application of Logistic Regression for Business Problems: An Overview
2. https://www.researchgate.net/publication/378754043_Understanding_Customer_Satisfaction_Factors_A_Logistic_Regression_Analysis - Understanding Customer Satisfaction Factors: A Logistic Regression Analysis

APPENDICES

Appendix A: Additional Visualizations







Appendix B: Model Code

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, roc_curve, auc, precision_recall_curve, average_precision_score

import seaborn as sns


# Load the datasets

training_data = pd.read_csv("C:/Users/reddy/Desktop/training.csv")
```

```
validation_data = pd.read_csv("C:/Users/reddy/Desktop/validation.csv")

# Handling missing values by imputing with median
median_income = training_data['Income'].median()
training_data['Income'].fillna(median_income, inplace=True)
validation_data['Income'].fillna(median_income, inplace=True)

# Feature selection
features = [
    'Year_Birth', 'Income', 'Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntFruits',
    'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds',
    'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
    'NumWebVisitsMonth'
]
target = 'Response'

# Preparing data for training and validation
X_train = training_data[features]
y_train = training_data[target]
X_validation = validation_data[features]
y_validation = validation_data[target]

# Data scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_validation_scaled = scaler.transform(X_validation)

# Logistic Regression Model
log_reg = LogisticRegression(max_iter=300)
log_reg.fit(X_train_scaled, y_train)

# Predictions
y_pred = log_reg.predict(X_validation_scaled)
y_pred_proba = log_reg.predict_proba(X_validation_scaled)[:, 1]
```

```
# Evaluation
accuracy = accuracy_score(y_validation, y_pred)
conf_matrix = confusion_matrix(y_validation, y_pred)
report = classification_report(y_validation, y_pred)

# Output results
print("Accuracy:", accuracy)
print("Confusion Matrix:\n", conf_matrix)
print("Classification Report:\n", report)

# Plotting Confusion Matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap='Blues', xticklabels=['Non-Response', 'Response'], yticklabels=['Non-Response', 'Response'])
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()

# ROC Curve
fpr, tpr, _ = roc_curve(y_validation, y_pred_proba)
roc_auc = auc(fpr, tpr)
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()

# Precision-Recall Curve
precision, recall, _ = precision_recall_curve(y_validation, y_pred_proba)
average_precision = average_precision_score(y_validation, y_pred_proba)
plt.figure(figsize=(8, 6))
```

```
plt.plot(recall, precision, color='blue', lw=2, label='Precision-Recall curve (AP = %0.2f)' % average_precision)
plt.fill_between(recall, precision, step='post', alpha=0.2, color='blue')
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.ylim([0.0, 1.05])
plt.xlim([0.0, 1.0])
plt.title('Precision-Recall Curve')
plt.legend(loc="lower left")
plt.show()
```