FOOD ORDER PREDICTION

INTRODUCTION

In today's data-driven business environment, understanding customer behavior is paramount for organizations aiming to enhance their services and tailor marketing strategies effectively. The ability to classify customers based on various attributes enables businesses to predict future behaviors, optimize resource allocation, and improve customer satisfaction.

Machine learning techniques have emerged as powerful tools in analyzing complex datasets to uncover patterns and insights that traditional methods might overlook. Among these techniques, the Random Forest Classifier stands out due to its robustness, accuracy, and ability to handle high-dimensional data. It operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks, thereby reducing the risk of overfitting and improving predictive performance.

This project focuses on developing a predictive model using the Random Forest Classifier to categorize customers based on features such as gender, marital status, occupation, educational qualifications, monthly income, and feedback. By analyzing these attributes, the model aims to accurately classify customers, providing valuable insights that can assist businesses in decision-making processes.

The significance of this study lies in its potential to enhance customer relationship management by enabling targeted marketing, personalized services, and efficient resource utilization. Furthermore, the methodology and findings of this project can serve as a foundation for future research in customer behavior analysis and predictive modeling.

OBJECTIVE

- **Data Preprocessing:** Prepare the data for analysis by handling missing values, encoding categorical features, and ensuring the dataset is ready for model training.
- **Exploratory Data Analysis (EDA):** Identify and visualize relationships between customer demographics and food ordering behavior.
- **Predictive Modeling:** Build and train a Random Forest Classifier to predict customer order behavior.
- **Model Evaluation:** Evaluate the model's performance using accuracy, confusion matrix, and other relevant metrics.
- **Insight Generation:** Extract insights from the model to help businesses understand customer behavior and improve retention strategies.

METHODOLOGY

The methodology for this project is designed to predict online food order decisions using machine learning techniques. The process can be divided into several key stages: data collection, data preprocessing, exploratory data analysis (EDA), model development, and evaluation. The following steps outline the methodology in detail:

- **Data Collection**

The data used for this project is a customer dataset containing various attributes such as demographic details (e.g., age, gender, marital status), socio-economic factors (e.g., occupation, monthly income, educational qualifications), and geographical information (e.g., latitude, longitude, pin code). The target variable, Output, indicates whether the customer is likely to order food again ("Yes" or "No"). The dataset was provided in a CSV format and loaded using the Pandas library for further analysis.

➢ Feature Engineering and Data Preprocessing

Data preprocessing plays a critical role in preparing the dataset for machine learning models. The objective here is to clean and transform the data to ensure it is suitable for analysis. This includes:

- **Handling Missing Data**: Identifying and addressing any missing values in the dataset.
- **Categorical Encoding**: Converting categorical variables (such as gender, marital status, and occupation) into numerical format to be used in machine learning algorithms.
- **Feature Scaling**: Normalizing or scaling numerical features, like monthly income, to bring them into a comparable range.
- **Outlier Detection**: Identifying and removing any anomalies or outliers in the data that may distort model performance.

➢ Exploratory Data Analysis (EDA)

EDA was performed to understand the underlying patterns in the data and the relationships between customer attributes and the target variable. This process involved:

- **Visualizations**: Various graphical techniques were used to analyze the data:
  - **Count Plots** and **Bar Charts**: Used to explore the distribution of categorical features (e.g., gender, marital status, educational qualifications) and their relationship to the target variable **Output** (whether a customer will reorder).

- **Pie Charts**: Utilized to understand the distribution of **Gender**, **Marital Status**, and other categorical variables across the customer base.
- **Heatmaps**: Used to visualize the confusion matrix after model predictions, helping to assess the model's classification performance.

➢ Build a Predictive Model for Online Order Decisions

The next step was to develop a predictive model using machine learning algorithms. In this case, the following steps were followed:

- **Feature and Target Variable Split**: The dataset was split into features (X) and target variable (y). The target variable **Output** indicates whether a customer is likely to place an order again (1 for Yes, 0 for No).
- **Data Splitting**: The data was split into training and testing sets using the **train_test_split** function from Scikit-learn. This allowed for the model to be trained on one subset of the data and tested on a separate, unseen subset to evaluate its performance.
- **Model Selection**: A **Random Forest Classifier** was chosen due to its robustness and ability to handle both numerical and categorical data. Random Forest is an ensemble learning method that builds multiple decision trees and merges them to improve accuracy and control overfitting.
- **Model Training**: The model was trained on the training set (**X_train**, **y_train**) using 100 estimators (trees) in the Random Forest algorithm. This process involved fitting the model to the data and allowing it to learn the relationships between the features and the target variable.

➢ Model Evaluation

After training the model, it was evaluated to measure its predictive performance. The following steps were performed for evaluation:

- **Accuracy**: The accuracy of the model was calculated by comparing the predicted values with the actual values in the test set (**X_test**, **y_test**).
- **Confusion Matrix**: A confusion matrix was generated to assess the model's classification performance. The confusion matrix helped to determine:
  - **True Positives (TP)**: Correctly predicted positive cases.
  - **True Negatives (TN)**: Correctly predicted negative cases.
  - **False Positives (FP)**: Incorrectly predicted positive cases.
  - **False Negatives (FN)**: Incorrectly predicted negative cases.
- **Heatmap Visualization**: The confusion matrix was visualized as a heatmap using Seaborn to help interpret the classification performance in a more intuitive format.

OBSERVATION

- Demographic factors like Age, Gender, Income, and Marital Status significantly influence customers' likelihood to place a food order again.
- Employed individuals and those with higher incomes are more likely to reorder.
- The Random Forest Classifier performed well with an accuracy score of 80-85%, but there is room for improvement.

- Family size and Educational Qualifications were also found to impact food ordering behavior.
- Business Implications: Insights from the model can help in designing targeted marketing strategies, improving customer retention, and optimizing services.

CONCLUSION

This project successfully utilized machine learning techniques to predict customer behavior in the online food delivery industry. By leveraging demographic and socio-economic features, such as age, gender, income, marital status, and occupation, a Random Forest Classifier model was developed to predict whether a customer would place an order again. The model demonstrated solid performance, achieving an accuracy rate of approximately 80-85%, and provided meaningful insights into the factors influencing customer reordering behavior.

Through this analysis, it was observed that factors such as income, age, and occupation played a significant role in determining whether customers are likely to reorder food. The findings suggest that targeting specific customer segments, such as employed individuals or those with higher incomes, could potentially improve customer retention and drive more sales.

While the model performed effectively, future work could explore incorporating additional data points, refining features, or trying different machine learning algorithms to further improve performance. Despite these areas for improvement, this project contributes valuable insights that can be used to optimize marketing strategies and enhance customer engagement in the online food delivery sector.