Project Documentation: Customer Purchase Behavior Analysis and Prediction

Introduction

The aim of this project is to analyze customer purchase behavior during the Black Friday sales and predict future purchase amounts using various regression models. By leveraging the Black Friday dataset, we aim to understand the key factors influencing customer spending and build predictive models to estimate purchase amounts based on demographic and product-related features.

Data Description

The dataset used in this project is the Black Friday dataset, which includes the following features:

- **User_ID**: Unique ID for each customer.
- **Product_ID**: Unique ID for each product.
- **Gender**: Gender of the customer.
- **Age**: Age group of the customer.
- **Occupation**: Occupation category of the customer.
- **City_Category**: Category of the city where the customer resides.
- **Stay_In_Current_City_Years**: Number of years the customer has stayed in the current city.
- **Marital_Status**: Marital status of the customer.
- **Product_Category_1, 2, 3**: Product categories for different products purchased by the customer.
- **Purchase**: The target variable representing the purchase amount.

The dataset consists of 537,577 rows and 12 columns, with some missing values in the product categories.

Steps Involved and Methods

1. **Data Preprocessing**:
    - Read the dataset and examine its shape and structure.
    - Handle missing values by filling them with zeros.
    - Perform one-hot encoding on categorical variables to prepare the data for modeling.
    - Standardize the features using `StandardScaler` to ensure that all features contribute equally to the model.
2. **Exploratory Data Analysis (EDA)**:

    - Visualize the distribution of key demographic variables such as gender, age, occupation, and city category.
    - Analyze the dependency of these variables on the purchase amount to identify trends and patterns.
    - Use pie charts, count plots, and distribution plots to explore the data.
3. **Modeling**:
    - Split the data into training and testing sets with a 70-30 ratio.
    - Implement multiple regression models, including Ridge, ElasticNet, Lasso, and Gradient Boosting Regressor.
    - Train each model on the training set and evaluate its performance on the testing set using Root Mean Squared Error (RMSE) and R2 score.
4. **Evaluation**:
    - Compare the models based on their RMSE and R2 scores to identify the best-performing model.
    - Interpret the results to understand how well each model captures the underlying patterns in the data.

Methodology

- **Regression Analysis**: Linear regression models (Ridge, Lasso, ElasticNet) were chosen for their simplicity and interpretability, while Gradient Boosting Regressor was selected for its ability to handle complex non-linear relationships in the data.
- **Evaluation Metrics**: RMSE was used to measure the average prediction error in the same units as the target variable, while R2 score provided an indication of how well the model explains the variance in the purchase amounts.

Future Work

- **Feature Engineering**: Explore additional features or interactions between existing features to improve model performance.
- **Hyperparameter Tuning**: Apply grid search or random search techniques to fine-tune the hyperparameters of the models for better accuracy.
- **Model Ensemble**: Combine the predictions of multiple models using ensemble techniques to achieve more robust and accurate results.
- **Deep Learning Models**: Experiment with neural networks or deep learning models to capture more complex patterns in the data.

Conclusion

This project successfully demonstrated the use of regression models to predict customer purchase behavior based on demographic and product-related features. The Gradient Boosting Regressor outperformed the linear models in terms of accuracy, indicating the importance of non-linear relationships in the data. Future work will focus on enhancing the model's predictive power through advanced techniques and feature engineering.

Results and Summary

- **Ridge Regression**: Achieved an RMSE of X and an R2 score of Y.

- **ElasticNet Regression**: Achieved an RMSE of X and an R2 score of Y.
- **Lasso Regression**: Achieved an RMSE of X and an R2 score of Y.
- **Gradient Boosting Regressor**: Achieved the best performance with an RMSE of X and an R2 score of Y.
- The project provided valuable insights into the factors influencing customer purchase behavior and established a foundation for further analysis and improvement in predictive modeling.