AD CLICK PREDICTION (Using Random Forest , Machine Learning)

Submitted By :
Aashima
(021323001)
Semester Four
Session : 2023 - 27

Bachelor Of Technology (B.Tech - AI ML)

Please feel free to reach out to me at the following
Email : aashimabhatia2005@gmail.com
Linkedin : https://www.linkedin.com/in/aashima-bhatia-a30919300/

College Dekho Assured(ImaginXp) Jagannath
Community College(Delhi)
Jagannath University,Bahadurgarh,Haryana

1 Introduction

This project develops a machine learning system to predict whether users will click on online advertisements. Using a dataset of 6,657 user interactions, it analyzes features like time spent, age, income, and ad topics to predict click behav- ior. Multiple models are compared, with the Random Forest classifier achieving the highest accuracy of 94.3%, supporting targeted advertising strategies.

2 KeyLibraries

- **Pandas**: Loads and manages the dataset (click_advertisement.csv).

- **NumPy**: Performs numerical operations for data preprocessing.

- **Scikit-learn**: Provides tools for data splitting, model training (Logistic Regression, Decision Tree, Random Forest), and evaluation.

- **XGBoost**: Implements the XGBoost classifier for comparison.

- **Seaborn/Matplotlib**: Visualizesdatadistributionsandmodelperformance.

- **Pickle**: Saves and loads the trained model for deployment.

3 KeyFeaturesoftheProject

1. **Click Prediction**: Predicts ad clicks based on user and ad features.

2. **Multiple Models**: Compares Logistic Regression, Decision Tree, Random Forest, and XGBoost.

3. **Feature Engineering**: Encodes categorical variables and scales numerical features.

4. **Large Dataset**: Processes 6,657 records with 14 features.

5. **Model Optimization**: Uses GridSearchCV to tune Random Forest hyperparameters.

6. **Evaluation Metrics**: Reports accuracy, confusion matrix, and classification report.

4 Core Algorithm

4.1 DataPreparation

- Loads dataset with 6,657 entries and 14 features (e.g., Time_Spent, Age, Ad_Topic).

- Encodes binary variables (Male, Clicked) using LabelEncoder.

• Convert categorical variables (Ad_Topic, $Country\_Name$, etc.) to one-hot encodings, yielding 304 features.

• Splits data into 70% training and 30% test sets.

• Scales numerical features using StandardScaler.

4.2 Model Architecture

• **Random Forest Classifier**: Ensemble of decision trees with 50 estimators, Gini criterion, and square root feature selection.

• **Other Models**: Logistic Regression, Decision Tree (entropy criterion), and XGBoost for comparison.

• Tunes Random Forest hyperparameters using GridSearchCV (n_estimators: 50, 100, 200; max_depth: None, 10, 20; min_samples_split: 2, 5, 10).

4.3 Training

• Trains models on scaled training data with random state for reproducibility

• Random Forest achieves 94.3% accuracy on test data after tuning.

4.4 Prediction and Evaluation

• Predicts clicks on test set using trained models. • Evaluates with accuracy,

confusion matrix, and classification report (precision, recall, F1-score).

• Random Forest confusion matrix: 1074 true negatives, 790 true positives, 31 false positives, 103 false negatives.

5 Conclusion

The project successfully predicts ad clicks using machine learning, with the Random Forest classifier achieving 94.3% accuracy. It outperforms Logistic Regression (89.9%), Decision Tree (91.0%), and XGBoost (94.0%) due to its ensemble approach. Limitations include high dimensionality from one-hot encoding and potential overfitting. Future work could explore feature selection, neural networks, or larger datasets for improved generalization.

6 Links

• **Demo Link**: https://eb386265c24c42ebf8.gradio.live

(Valid only for one week)

IMAGES :