### **Predicting Ads Click-Through Rate: A Case Study in Classification on Imbalanced Data**

### This project focuses on predicting the Click-Through Rate (CTR) of online advertisements using a dataset containing information about ad impressions, user interactions, and ad features. The primary goal is to build a robust classification model that can accurately identify instances where users are likely to click on ads, even in the face of imbalanced data. This project demonstrates the process of handling class imbalance in classification, highlighting the challenges and solutions involved.

### **Project Goal**

### The main goal is to develop a machine learning model capable of predicting whether a user will click on an advertisement. This prediction is crucial for advertisers to optimize their campaigns, improve user targeting, and ultimately increase return on investment (ROI).

### **Project Process**

### **Data Loading and Exploration:** Loaded the ads dataset using pandas and explored its structure, data types, and missing values.

### **Data Visualization:** Created visualizations to understand the distribution of click status (clicked vs. not clicked), numerical features (user age, ad duration, impression count), and categorical features (region, ad type, device type).

### **Data Preprocessing:** Handled class imbalance by oversampling the minority class (ads clicked) using the resample method from scikit-learn.

### **Feature Engineering:** Encoded categorical features using LabelEncoder and created new features based on existing data to improve model performance.

### **Model Training:** Trained a Random Forest classifier on the oversampled data to predict click status.

### **Model Evaluation:** Evaluated the model's performance using metrics like the classification report, accuracy score, precision, recall, F1-score, and AUC-ROC.

### **Comparison with Original Data:** Applied the trained model to the original dataset and compared predictions with actual click statuses.

### **Visualization of Results:** Created a bar chart to visualize the classification accuracy and feature importance.

### **Methods**

### ***Data Exploration and Preprocessing: The project starts with an in-depth analysis of the dataset, including:***

### Examining the distribution of classes in the target variable (click\_status) to understand the level of imbalance.

### Visualizing the distributions of key numerical features like ad duration, impression count, and user demographics.

### Analyzing the distribution and relationships of categorical features such as region, ad type, and device type.

### **Handling Class Imbalance:** Since the dataset exhibits a significant imbalance between the click\_status classes (more instances of non-clicks than clicks), the project utilizes oversampling to balance the classes. This involves replicating instances from the minority class to achieve an equal representation of both classes.

### **Feature Selection:** To identify the most impactful features for prediction, the project employs feature importance techniques. This involves training a Random Forest model and analyzing the feature importance scores to identify the top variables influencing the click\_status prediction.

### **Model Training:** Using the balanced dataset and selected features, a Random Forest classifier is trained to predict the click\_status.

### **Model Evaluation:** The trained model is evaluated on the test dataset using metrics suitable for imbalanced datasets, such as:

### **Precision:** Proportion of correctly predicted positive instances out of all predicted positives.

### **Recall:** Proportion of correctly predicted positive instances out of all actual positives.

### **F1-Score:** Harmonic mean of precision and recall.

### **Accuracy:** Overall proportion of correctly classified instances.

### **Results**

### The developed Random Forest model achieved impressive performance on the test dataset, demonstrating high precision, recall, and F1-score for both classes. The high recall for the "clicked" class is particularly noteworthy, indicating that the model is effective at identifying instances where users are likely to click on ads.

### **Features**

### Key Features: The project identified features such as ad duration, user age, impression count, region, ad type, and device type as having a significant influence on predicting click-through rates.

### Unique Identifier: While ad\_id was initially identified as a highly important feature, it was removed during model training as it is not directly relevant for prediction.

### **Tools**

### Python: The core programming language for data processing, analysis, and model building.

### Pandas: A powerful library for data manipulation and analysis.

### Scikit-learn (sklearn): A widely used library for machine learning algorithms, including Random Forest classifiers.

### Imbalanced-learn (imblearn): A library specifically designed for handling imbalanced datasets, including oversampling techniques like SMOTE.

### Matplotlib and Seaborn: Libraries for data visualization.

### **Importance**

### This project has significant implications for the digital advertising industry:

### Campaign Optimization: The model can help advertisers better assess which ads are likely to engage users, leading to more effective ad placements and increased click-through rates.

### Cost Efficiency: By identifying patterns associated with user interactions, the model can assist advertisers in allocating their budgets more effectively.

### User Targeting: The model can be used to identify user segments with higher engagement, allowing advertisers to develop targeted advertising strategies.

### **Conclusion**

### This project showcases the process of building a robust classification model for predicting ads click-through rates, addressing the challenges associated with imbalanced data. By utilizing appropriate methods for handling class imbalance, feature selection, and model evaluation, the project demonstrates the potential of machine learning to improve advertising strategies and decision-making in the digital marketing industry.AUC (Area Under the ROC Curve): Measures the overall ability of the model to distinguish between the two classes.

### **References**

* Ads Click Through Rate Prediction using Python: Aman KharwalKharwal & Lemartinezba<https://thecleverprogrammer.com/2023/01/16/ads-click-through-rate-prediction-using-python/>