Capstone Project Proposal - Click-Through Rate (CTR) Prediction
Name: Vidushi Raval

Problem Identification:
This project focuses on predicting whether a user will click on an advertisement by analyzing factors such as user demographics, ad attributes, and contextual information.
Accurate click-through rate (CTR) prediction is essential for digital advertising platforms like Google Ads, Meta Ads, and TikTok, enabling them to optimize ad placements, refine targeting strategies, and maximize advertiser revenue.
By leveraging machine learning techniques, this project aims to develop a model that effectively estimates the probability of a user engaging with an ad, ultimately improving campaign effectiveness and user experience.
The project aims to predict the likelihood of a user clicking on an ad based on various factors like user demographics, ad attributes, and contextual information.

Problem Statement Formation:
The objective is to make online ads more effective by improving CTR prediction accuracy. This will help advertisers spend their budgets wisely and show users ads that are more relevant to them.
Using machine learning, we'll build a model that performs better than basic approaches and provides valuable insights for the digital advertising industry.

Context:
In digital advertising, predicting CTR helps platforms maximize revenue by displaying ads to users who are more likely to engage.
This enables advertisers to allocate budgets more efficiently and improves user experience by displaying relevant ads.
The Avazu CTR Prediction Dataset from Kaggle contains anonymized user and ad interaction data, providing a rich source of information to build predictive models.

Goal of the Project:
1) Increase Ads conversion revenue: Develop a machine learning model that accurately predicts the likelihood of a user clicking on an ad based on various features such as user demographics, ad details, and context.
2) Optimize Ads spending: Use the predicted CTR to optimize ad placement, enhance targeting strategies, and improve overall ad campaign effectiveness.

Criteria for success:
- Develop a machine learning model that accurately predicts the probability of a user clicking on an ad.
- Improve the model's performance such as a logistic regression model with default settings, and demonstrate enhancements compared to baseline approaches like logistic

regression or random prediction.
- Provide meaningful insights into feature importance and their impact on CTR.
- Create a well-organized GitHub repository with clear documentation, including code, a detailed project report, and a presentation deck to showcase findings.

Scope of solution space:

1. Extract relevant features from categorical and timestamp variables.
2. **Data Preprocessing**
    - Handling missing values and outliers
    - Encoding categorical variables
    - Feature scaling and transformation
3. **Feature Engineering**
    - Extracting meaningful features from timestamp data (e.g., time of day, day of the week)
    - Creating user-device interaction features
    - Identifying important ad placement factors
4. **Model Selection**
    - Training and evaluating multiple machine learning models, such as:
        - Logistic Regression
        - Decision Trees
        - Random Forest
5. **Cross-validation**
   Cross-validation ensures that the model's performance is reliable across different subsets of data.
6. **Interpretability & Insights**
    - Providing actionable insights for advertisers and ad platforms

Constraints:

- Large dataset size (~40 million records) may require computational optimizations.
- Data is anonymized, limiting the ability to derive deep behavioral insights.
- Address the issue of significantly more non-clicks than clicks.
- Need for effective feature selection and engineering to avoid overfitting.

Stakeholders:

- **Digital advertisers:** Businesses running ad campaigns who benefit from better CTR predictions.
- **Ad networks:** Platforms like Google Ads, Meta Ads, and TikTok optimizing ad placements.
- **Data scientists & engineers:** Professionals working on machine learning models for ad-tech.
- **End users:** Consumers who interact with advertisements online.

Data sources:

The Avazu Click-Through Rate (CTR) Prediction Dataset on Kaggle contains millions of

records with information such as user demographics, ad details, and contextual information like device type and timestamp.
This data will be used to train machine learning models to predict the probability of a user clicking on a given ad.

Dataset: https://www.kaggle.com/c/avazu-ctr-prediction
There are three types of files.

- **train** - Training set. 10 days of click-through data, ordered chronologically. Non-clicks and clicks are subsampled according to different strategies.
- **test** - Test set. 1 day of ads for testing your model predictions.

Available Data fields:

- id: ad identifier
- click: 0/1 for non-click/click
- hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
- C1 -- anonymized categorical variable
- banner_pos
- site_id
- site_domain
- site_category
- app_id
- app_domain
- app_category
- device_id
- device_ip
- device_model
- device_type
- device_conn_type
- C14-C21 -- anonymized categorical variables

This dataset can be used to build a model that predicts the probability of a user clicking on an ad, which is crucial for ad optimization and targeting in digital advertising platforms.

Approach:

- **Data Exploration**:
    - Understand the dataset distribution,
    - missing values, and correlations.
- **Preprocessing**:
    - Clean and transform data,
    - Encode categorical variables,
    - Create new features.

- **Feature Engineering:**
    - Identify high-impact features using exploratory data analysis (EDA).
    - Reduce dimensionality for high-cardinality categorical variables.
- **Model Training**:
    - Train baseline models and experiment with advanced algorithms.
    - Experiment with models such as Logistic Regression, Decision Trees and Random Forest.
    - Compare models and accuracy.
- **Interpretability & Insights:**
    - Analyze feature importance and interpret results.
    - Provide actionable recommendations for optimizing ad placements.
- **Evaluation**: Compare models using AUC-ROC(Area Under the Receiver Operating Characteristic Curve) to measure how well a machine learning model can distinguish between positive and negative classes. Precision-recall, and log loss.
- **Final Deliverables:**
    - A GitHub repository containing all code for data processing, modeling, and evaluation, project files and documentation.
    - A project report summarizing findings, methodology, and conclusions.
    - A slide deck presentation for stakeholders.

By implementing this project, we aim to develop a robust CTR prediction model that enhances digital advertising efficiency and effectiveness.
This project aims to provide valuable insights into CTR prediction, demonstrating how machine learning can enhance digital advertising strategies.