Project Proposal: Evergreen or Ephemeral?

Question:

The question stems from a Kaggle challenge posed by StumbleUpon, a user-curated web content discovery engine, which recommends pages and media to its users.

On the site, users can rate whether a recommended page is “evergreen” or “ephemeral”. StumbleUpon wants to improve its ability to recommend relevant content by identifying up-front whether a web page is likely to be “evergreen”.

The goal is to build a classifier that would determine whether a web page is “evergreen” in order to improve StumbleUpon’s recommendation system.

Data:

The data consists of 7000+ rows, with one row of data corresponding to data points related to one specific URL. These data points include but are not limited to category, spelling errors ratio and more.

The target is evergreen consists of two outcomes (1) if yes and (0) if no.

Tools:

- Pandas and Numpy for Data Cleaning and Exploration
- Matplotlib and Seaborn for Visualization
- ScikitLearn for modeling: knn, logistic regression, random forest xgboost

MVP Goal:

- Baseline modeling with Log Reg and KNN
- Class imbalance treatment if needed