**UDACITY - Machine Learning Engineer Nanodegree**

**Capstone Project Proposal**

Connor Phoenix

12.6.2021

**Project** - Artist Recommendation (via Sagemaker Factorization Machine)

**Domain** - Content recommendation is one of the most popular machine learning applications. Essential to the business model of many modern companies is the ability to offer users tailored content based on available user engagement data. This challenge is ubiquitous across many domains such as e-commerce and audio/video streaming services. Effectively solving the content recommendation problem was a key success factor for companies such as Amazon and Netflix. Interestingly the maturation of MLaaS platforms like AWS Sagemaker has lowered the overhead in developing and deploying such systems.

**Problem Statement** - Given a dataset of artist play counts by user, create a system to generate an array of artist recommendations for any input user.

**Dataset** - The dataset was sourced from the Last.fm API by Oscar Celma of Pompeu Fabra University. The data is available for non-commercial public use via the university website:

- https://www.upf.edu/web/mtg/lastfm360k

An ingestion and initial preprocessing script can be found in the Udacity_ML_Recommendation repository.

**Solution Framework**

1. Merge artist play and user files into single dataset
2. Sample data by x% (% to be determined)*
3. Perform some exploratory data analysis
4. Filter to top n% of artists (% to be determined)*
5. Convert user/artist plays into binary target variable:
    For every user, percentile rank artist play counts
    For artists in quartiles 3/4 label 1
    For artists in quartiles 1/2 label 0
6. Create a sparse matrix where rows represent users with the following features:
    Binary 1/0 variables for every artist in sampled/filtered dataset
    Additional user features (i.e. age, gender, country etc.)
7. Use sparse matrix to generate training/test datasets
8. Fit a Sagemaker “factorization machine” estimator
9. Perform hyperparameter tuning
10. Use predict probabilities to generate ranked artist recommendations for any user.

*Sampling needed due to resource constraints (raw dataset has 17M records). Potential future work could be to expand the approach to the entire dataset using larger instance types and/or distributed processing.

**Many artists in the dataset have very low play counts which could introduce noise into the

system. The idea is that by filtering to the top artists we are ensuring a certain level of quality in the generated recommendations. This also has the side benefit of further reducing the dataset size.

**Evaluation** - Per Sagemaker Factorization Machine Documentation, model tuning can optimize for accuracy, cross entropy, or beta. Additionally, I will evaluate the classifier using AUC (Area Under ROC curve). Beyond quantitative evaluation, spot checking specific user recommendations will also help provide intuition on how the system is performing.

**Benchmarking** - While I could not find a case study for this exact dataset and problem statement the below Kaggle competition will serve as a suitable proxy. Submissions were evaluated based on AUC (Area Under ROC curve).

**https://www.kaggle.com/c/kkbox-music-recommendation-challenge/leaderboard**